Spring 5-2020

Phytoplankton Community Composition in the Surface Ocean: Methods for Detection using Optical Measurements, Pigment Concentrations, and Flow Cytometry

Alison P. Chase
University of Maine, alisonpchase@gmail.com

PHYTOPLANKTON COMMUNITY COMPOSITION IN THE SURFACE OCEAN: METHODS FOR DETECTION USING OPTICAL MEASUREMENTS, PIGMENT CONCENTRATIONS, AND FLOW CYTOMETRY

By

Alison P. Chase

B.A. Bowdoin College, 2009

M.S. University of Maine, 2014

A DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

(in Oceanography)

The Graduate School

The University of Maine

May 2020

Advisory Committee:

Emmanuel Boss, Professor of Marine Sciences, Co-advisor

Lee Karp-Boss, Associate Professor of Marine Sciences, Co-advisor

Collin Roesler, Professor of Earth and Oceanographic Science, Bowdoin College

Heidi Sosik, Senior Scientist, Woods Hole Oceanographic Institution

Lawrence Mayer, Professor of Oceanography

PHYTOPLANKTON COMMUNITY COMPOSITION IN THE SURFACE OCEAN: METHODS FOR DETECTION USING OPTICAL MEASUREMENTS, PIGMENT CONCENTRATIONS, AND FLOW CYTOMETRY

By Alison P. Chase

Dissertation Advisors: Dr. Emmanuel Boss, Dr. Lee Karp-Boss

An Abstract of the Dissertation Presented
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
(in Oceanography)
May 2020

Phytoplankton are microscopic photoautotrophs living in the surface ocean waters and help support all life on earth via photosynthetic production of oxygen. Thousands of species make up the bulk phytoplankton community, and the spatial and temporal distribution of different types of phytoplankton has relevance for many ocean ecosystem processes including marine food web dynamics, and carbon flux and sequestration. Earth-observing satellites offer a method to detect phytoplankton community composition (PCC) on the vast scale of the global ocean. The use of satellite data to observe and interpret PCC in the surface ocean requires significant effort to develop and evaluate algorithms based on measurements made in situ; the work of this thesis contributes to that effort.

Information from both global and regional (North Atlantic Ocean) datasets is applied to develop methods to estimate phytoplankton pigment concentrations, phytoplankton size classes, and diatom carbon concentrations. Optical spectra, specifically hyperspectral remote-sensing reflectance, are used in the algorithm for estimating phytoplankton pigments, which resolves the concentrations of three pigments and one pigment group (chlorophylls *a, b, c,* and photoprotective carotenoids). This has implications for use with hyperspectral ocean color data measured by satellite. A novel dataset of open-ocean imaging-in-flow cytometry is used to evaluate and improve a commonly applied phytoplankton size class algorithm, as well as to calculate diatom carbon and develop a model to map diatom carbon

using environmental parameters as model input. Biases and uncertainties in the size class algorithm are reduced by our method relative to previously published work for all three size classes (pico-, nano-, and microplankton). Diatom carbon measurements from quantitative cell imagery elucidate the variability of diatom biomass as function of chlorophyll *a* concentration, and this information enables development of a novel approach to detect diatom carbon from space.

The findings of this thesis are relevant to large-scale studies of ocean ecosystems and are critical for algorithm development using both current and upcoming earth-observing satellite data. Additionally, the results presented here provide tools that will benefit oceanographic research on spatial scales relevant to a changing ocean climate.

DEDICATION

To the Ocean.

As humans, we cannot reasonably hope to ever fully comprehend and understand the vast size and complexity of this amazing body of water and all the life within it, but I am extremely grateful for the opportunity to spend a significant part of my life trying.

ACKNOWLEDGEMENTS

The science presented in this thesis would not have been possible without the support of many people, from whom I learned much about oceanography, life, and everything in between. Without a doubt I have Emmanuel Boss and Lee Karp-Boss to thank first. They have seen me through growth, both personal and professional, and I could not be more grateful to count my advisors as not only colleagues, but also as friends, especially in a field of research that is more of a lifestyle than a job. Thank you for your passion and enthusiasm, for sharing your knowledge, and for encouraging me to solve problems on my own and in doing so grow towards becoming an independent researcher.

Thanks to my colleagues who are also some of my dearest friends, who work hard, and most importantly, make science fun: Sasha, Jason, Paula, Ivona, Jeremy, Wayne, Cael, Kelsey, Andre, Guillaume, and others. To the officemates who kept me company in Aubert 461B over the years and who became great friends: Karen, Kerstin, Nick, Lino, Jordan, Juliana, and especially Nils – work (and life) just won't be the same without you! Many thanks to my colleagues in the (sub)mesoscale group at UW: Pete, Kyla, AliD, Emmett and the rest – it is an honor to work with such a wonderful group of excellent scientists and truly kind and caring people.

I am very appreciative of my Master's and Doctoral committee members who contributed their time and expertise to helping me become a better scientist: Mary Jane Perry, Annick Bricaud, Collin Roesler, Larry Mayer, and Heidi Sosik. To Collin especially, who helped launch me on my journey towards becoming an oceanographer when I was an undergraduate student, and who has pushed me to become the best I can be and whose guidance I am extremely grateful for.

Thanks to the co-authors of the manuscript chapters of this thesis for all your help and support; Chapter 3: Emmanuel Boss, Ivona Cetinic, and Wayne Slade; Chapter 4: Sasha Kramer, Nils Haëntjens, Emmanuel Boss, Lee Karp-Boss, Mimi Edmondson, and Jason Graff; Chapter 5: Emmanuel Boss, Lee

Karp-Boss, Nils Haëntjens, and Emmett Culhane. Thanks to Mike B and the NAAMES family – my life is forever changed. Thank you to the University of Maine faculty and staff – Wge, Andy, Maura, Sue, Jodie, Jessie, Carrie, and others – whose friendship and help is much appreciated. To our community and friends in Orono who made our years here a chapter we are so grateful for, thanks. To Jim Dale, whose voice provided soothing background noise to many hours of work. Thanks to my family for all the love and opportunities. To my favorite RDF, thank you, for everything. To Ken, who has been by my side since the start of this project: thank you, so much.

TABLE OF CONTENTS

DEDICATION iii

ACKNOWLEDGEMENTS iv

LIST OF TABLES x

LIST OF FIGURES xii

LIST OF EQUATIONS xv

LIST OF ABBREVIATIONS xviii

Chapter

1. INTRODUCTION 1

1.1 Background 1

1.2 Defining Phytoplankton Community Composition 2

1.3 Measuring PCC on Global and Ocean Basin Scales 5

1.4 Thesis Data and Objectives 9

2. PERSPECTIVES ON HYPERSPECTRAL OPTICAL MEASUREMENTS 13

2.1. Why Measure Optical Properties? 13

2.2. Why Hyperspectral? 14

2.3. Deriving Information from Hyperspectral Measurements 15

2.3.1. Direct Use of Optical Spectra 16

2.3.2. Methods of Spectral Inversion 19

2.4. Making the Leap to Space-based Measurements 23

2.5. Take-home Points 25

3. ESTIMATION OF PHYTOPLANKTON ACCESSORY PIGMENTS FROM HYPERSPECTRAL REFLECTANCE SPECTRA: TOWARD A GLOBAL ALGORITHM 26

3.1 Introduction 26

3.2 Data and Methods 30

3.2.1 Data Sets 30

3.2.1.1. Tara Expeditions 30

3.2.1.2 SABOR, A1319, and NH1418 Expeditions 31

3.2.2. $R_{rs}(\lambda)$ Data and Processing 34

3.2.2.1. Raman Scattering Correction 37

3.2.2.2. Correction for Angular Effects 37

3.2.2.3. Inversion of $R_{rs}(\lambda)$ Spectra 38

3.2.3. Estimation of Pigment Concentrations 41

3.2.3.1. Pigments Estimated from $a_{gaus}(\lambda)$ –HPLC Relationships 44

3.2.3.2. Pigments Estimated from TChl a_{OC4} Using HPLC Covariation Relationships 46

3.2.3.3. Calculation of Uncertainties in *A* and *B* Coefficients and in Estimated Pigments 48

3.2.4. Analysis of Spectral Residuals 49

3.2.4.1. Calculation of $a_{p\text{-}global}(\lambda)$ 49

3.2.4.2. Calculation an Inversion of Spectral Residuals 52

3.3. Results 52

3.3.1. Pigments Estimated from $a_{gaus}(\lambda)$ and Covariation Relationships 52

3.3.2. Pigment Ratios 56

3.3.3. Spectral Residuals 57

3.4. Discussion 57

3.4.1. Estimation of Pigment Concentrations 58

3.4.2. Pigment Covariation 59
3.4.3. Spectral Resolution Considerations 60
3.4.4. Spectral Residuals Analysis 61
3.5. Conclusions 62
3.6. Comparison of Chlorophyll Algorithms 63
3.7. Supplementary Information 64
4. EVALUTION OF DIAGNOSTIC PIGMENTS TO ESTIMATE PHYTOPLANKTON SIZE CLASSES 69
4.1 Introduction 69
4.2 Materials and Procedures 73
4.2.1 HPLC Pigment Data and Application of the DPA 74
4.2.2. Collection and Analysis of Cytometry Data 76
4.2.3. Biovolume Estimates of Pico- Nano- and Microplankton from Merged Cytometry Data 78
4.2.4. Statistics Used for Method Evaluation 81
4.3. Assessment and Discussion 82
4.3.1. Overall Trends in PSC Estimate Comparisons 82
4.3.2. Discrepancies Between Attribution of Pigments to PSCs, and Observed Cell Sizes and Groups 85
4.4. Recommendations 89
5. DIATOM CARBON FROM SPACE: A MODEL FOR THE WESTERN NORTH ATLANTIC OCEAN 94
5.1. Introduction 94
5.2. Study Region and Datasets 97
5.2.1. Imaging-in-flow Cytometry 98
5.2.2. Phytoplankton Classification and Carbon Estimates 99

5.2.3. HPLC Pigments and CHEMTAX 100
5.2.4. Pigments Estimated from Particulate Absorption Spectra ($a_p(\lambda)$) 103
5.2.5. Satellite Data 103
5.3. Approach 104
5.3.1. Diatom Carbon Neural Network Setup and Testing 105
5.3.2. Evaluating Model Uncertainties 107
5.4. Results and Discussion 110
5.4.1. Diatoms Estimated from HPLC Pigments 110
5.4.2. Diatom Variability as a Function of Chl *a* 111
5.4.3. Satellite-based Diatom Carbon 114
5.5. Conclusions 117
6. CONCLUSIONS AND FUTURE DIRECTIONS 118
6.1. Future Work 120
6.2. Summary 121
REFERENCES 123
APPENDIX: Supplementary Information to Chapter 5 141
BIOGRAPHY OF THE AUTHOR 162

LIST OF TABLES

Table 1.1. Taxonomic classification and size classes of common oceanic phytoplankton 4

Table 1.2. Occurrence of pigments within phytoplankton groups 8

Table 2.1. Summary of previous studies estimating phytoplankton groups or pigments from hyperspectral optical measurements.....22

Table 3.1. Summary of datasets used in the present chapter.....31

Table 3.2. Pigment groups from HPLC analysis 32

Table 3.3. Parameters allowed to vary during the inversion of $R_{rs}(\lambda)$ spectra.....42

Table 3.4. Coefficients of $a_{gaus}(\lambda)$/HPLC relationships, and HPLC co-variation relationships 48

Table 3.5. Statistics of pigment estimation 54

Table 3.6. Statistics calculated from the comparison of the OC algorithms with HPLC TChl *a* 64

Table 3.7. Spectral coefficients and errors calculated for the Chl *a* and $a_p(\lambda)$ global relationship.....65

Table 4.1. Formulas to calculate picoplankton (*Fp*), nanoplankton (*Fn*), and microplankton (*Fm*) 75

Table 4.2. Statistics of PSC fractions estimated using published versions of the DPA compared to cytometry data 84

Table 4.3. Statistics of PSC Chl *a* concentrations estimated using published versions of the DPA compared to cytometry data.....84

Table 4.4. Assignments of the seven accessory pigments used in the DPA to phytoplankton groups and size classes.....85

Table 4.5. Statistics of PSC fractions and Chl *a* concentrations estimated using updated DPA equations compared to cytometry data 93

Table 6.1. Summary of data types, approaches, and associated products 122

Table A.1. Image categories used for input to the Random Forest machine learning built into the EcoTaxa application..... 142

Table A.2. Image categories assigned via manual correction and confirmation of images on EcoTaxa..... 145

LIST OF FIGURES

Figure 1.1. Example spatial resolution of different data types 6

Figure 1.2. Relative biovolume of major identifiable phytoplankton groups 10

Figure 2.1. Pigment-specific absorption spectra ($a^*(\lambda)$) across the visible spectrum 16

Figure 2.2. Specific absorption spectra of three representative phytoplankton phyla 19

Figure 2.3. Example particulate absorption ($a_p(\lambda)$, m^{-1}) spectrum with component Gaussian functions 21

Figure 2.4. Median reflectance spectrum taken in the Azov Sea with marked locations of spectral bands of MERIS (A) and the first derivative of reflectance (B) 24

Figure 3.1 Locations of all data used in the present chapter 33

Figure 3.2 $R_{rs}(\lambda)$ spectra measured in situ with a Satlantic HyperTSRB 36

Figure 3.3 Schematic of the data and processing steps used to estimate accessory pigments from hyperspectral $R_{rs}(\lambda)$ data 44

Figure 3.4 Comparisons between HPLC pigment concentrations and $a_{gaus}(\lambda)$ magnitudes 46

Figure 3.5 Relationships between accessory pigments and TChl *a* 47

Figure 3.6 Chl *a* estimated using a_p(676 nm) line height regressed against in situ-measured a_p(440 nm) 51

Figure 3.7. Spectral A_{chl} and B_{chl} derived from the relationship: $a_p(\lambda) = A_{chl} * \text{Chl } a^{B_{chl}}$ 51

Figure 3.8. Comparisons between HPLC pigment concentrations and pigments 55

Figure 3.9. Comparison of TChl *a* computed from spectra with HPLC TChl *a* 55

Figure 3.10. Frequency distributions of pigment ratios 56

Figure 3.11. Chl *a* estimated from $R_{rs}(\lambda)$ spectra 64

Figure 4.1. Data locations, temperature and salinity values, and Chl *a* concentrations.....74

Figure 4.2. Cell biovolume and Chl *a* compared to Chl *a* from HPLC 80

Figure 4.3. Comparison of size classes estimated by cytometry and pigments.....83

Figure 4.4. Summed biovolume of phytoplankton cells imaged by the IFCB across all samples. 87

Figure 4.5. Comparison of size classes estimated by cytometry and pigments using updated equations. 92

Figure 5.1. Locations of the four NAAMES cruise tracks in the western North Atlantic Ocean. 98

Figure 5.2. Satellite data used for neural network model input 104

Figure 5.3. Diatom carbon (mg m^{-3}) estimated from cell imagery; n = 1,449 107

Figure 5.4. Flow chart of neural network model training and application.....107

Figure 5.5. Results of neural network training and testing.....109

Figure 5.6. Frequency distribution of neural network model error.....109

Figure 5.7. Diatom carbon estimated from IFCB imagery and from the Diagnostic Pigment Analysis (DPA) following H11.....111

Figure 5.8. Diatom carbon estimated from IFCB imagery and using the formula of H11 compared to Chl *a*.....112

Figure 5.9. Correlations between diatom carbon and environmental parameters across four Chl *a* concentrations.....113

Figure 5.10. Diatom carbon estimated from satellite data for the northwest Atlantic Ocean in May 2016 115

Figure 5.11. Percent difference in diatom carbon 116

Figure A.1. The number of IFCB samples combined as a function of Chl *a* concentration.....152

Figure A.2. The number of IFCB samples combined to decrease statistical counting errors in IFCB data 153
Figure A.3. Number of diatoms per IFCB sample 153
Figure A.4. Phytoplankton group contribution to Chl *a* from CHEMTAX analysis with inputs from Swan et al. 2016 156
Figure A.5. Phytoplankton group contribution to Chl *a* from CHEMTAX analysis with inputs from van de Poll et al. 2013 156
Figure A.6. Diatom contribution to Chl *a* from IFCB imagery 157
Figure A.7. Diatom carbon from IFCB imagery vs. diatom carbon from CHEMTAX 157
Figure A.8. Diatom Chl *a* from the DPA and from CHEMTAX 158
Figure A.9. Locations and frequency distributions of the five input parameters used for neural network training 159
Figure A.10. Diatom carbon from IFCB imagery as a function of environmental parameters and pigment anomalies 160
Figure A.11. Chl *a* concentration from HPLC pigment analysis vs. from Gaussian decomposition of $a_p(\lambda)$ 161
Figure A.12. Chl *a* vs. fucoxanthin as determined from HPLC pigment analysis 161

LIST OF EQUATIONS

2.1. $a(\lambda) = a_{\varphi}(\lambda) + a_{CDOM}(\lambda) + a_{NAP}(\lambda) + a_w(\lambda)$ 15

2.2. $R_{rs}(\lambda) = \frac{f}{Q}\frac{b_b(\lambda)}{a(\lambda)+b_b(\lambda)}$ 15

3.1. $R_{rs}(\lambda) = \frac{f}{Q}\frac{b_b(\lambda)}{a(\lambda)+b_b(\lambda)}$ 27

3.2. $R_{rs}(\lambda) = \frac{L_w(\lambda,0^+)}{E_d(\lambda)}$ 34

3.3. $L_u(\lambda, z_{0-}) = L_u(\lambda)e^{-K_{Lu}(\lambda)*z}$ 34

3.4. $K_{Lu}(\lambda) \cong \frac{a_w(\lambda)+a_p(\lambda)}{\mu_u}$ 34

3.5. $L_w(\lambda, 0^+) = \frac{t\, L_u(\lambda,z_{0-})}{n^2}$ 36

3.6. $u(\lambda) \equiv \frac{b_b(\lambda)}{a(\lambda)+b_b(\lambda)}$ 38

3.7. $r_{rs}(\lambda) = g_1 u(\lambda) + g_2 u(\lambda)^2$ 38

3.8. $r_{rs}(\lambda) = \frac{R_{rs}(\lambda)}{0.52+1.7R_{rs}(\lambda)}$ 38

3.9. $\chi^2 = \sum_{i=1}^{60}\left(\frac{u_{\text{meas}}(\lambda_i)-u_{\text{mod}}(\lambda_i)}{u_{\text{std}}(\lambda_i)}\right)^2$ 39

3.10. $u_{\text{mod}}(\lambda) = \frac{b_{bp}(\lambda)+b_{bw}(\lambda)}{a_{\varphi}(\lambda)+a_{CDOM}(\lambda)+a_{NAP}(\lambda)+a_w(\lambda)+b_{bp}(\lambda)+b_{bw}(\lambda)}$ 39

3.11. $b_{\text{bp}}(\lambda) = \frac{b_{\text{bp}}}{b_{\text{p}}}\left(c_{\text{p}}(\lambda) - a_{\text{p}}(\lambda)\right)$ 39

3.12. $c_{\text{p}}(\lambda) = C_{cp}\left(\frac{\lambda}{\lambda_0}\right)^{-\gamma}$ 39

3.13. $a_{\text{NAP}}(\lambda) = C_{\text{NAP}}e^{-S_{NAP}(\lambda-\lambda_0)}$ 40

3.14. $a_{\text{CDOM}}(\lambda) = C_{\text{CDOM}}e^{-S_{CDOM}(\lambda-\lambda_0)}$ 40

3.15. $a_{\varphi}(\lambda) = \sum_{i=1}^{8} a_{gaus}(peak_i, \lambda)e^{\left(-0.5\left(\frac{\lambda-peak_i}{\sigma_i}\right)\right)^2}$ 40

3.16. $a_{gaus}(\lambda) = A_{inv}[\text{pigment}_{\text{HPLC}}]^{B_{inv}}$ 44

3.17. $\chi^2 = \sum_{i=1}^{97}\left(a_{gaus}(\lambda)_i - A_{inv} * [\text{pigment}_{\text{HPLC}}]_i^{B_{inv}}\right)^2$ 45

3.18. $[\text{pigment}_{\text{inv}}] = \left(\frac{a_{gaus}}{A_{inv}}\right)^{\frac{1}{B_{inv}}}$ 45

3.19. $ME = \left(\text{median}\left(\frac{\text{abs}([\text{pigment}_{\text{inv}}]-[\text{pigment}_{\text{HPLC}}])}{[\text{pigment}_{\text{HPLC}}]}\right)\right) * 100$ 45

3.20. $\log(\text{TChl } a) = \log A_{cov} + B_{cov} * \log([\text{pigment}_{\text{HPLC}}])$ 47

3.21. $\text{TChl } a = A_{cov}[\text{pigment}_{\text{HPLC}}]^{B_{cov}}$ 47

3.22. $[\text{pigment}_{\text{cov}}] = \left(\frac{\text{TChl } a_{\text{OC4}}}{A_{cov}}\right)^{\frac{1}{B_{cov}}}$ 47

3.23. $\chi^2 = \sum_{\lambda=400}^{700}\left(\frac{a_p(\lambda)-A_{chl}(\lambda)Chl^{Bchl(\lambda)}}{a_{pUNC}(\lambda)}\right)^2$ 50

3.24. $a_{\text{p-global}}(\lambda) = A_{chl}(\lambda)TChl\ a^{Bchl(\lambda)}$ 50

3.25. $u_{\text{resid}}(\lambda) = u_{\text{meas}}(\lambda) - u_{\text{mod}}(\lambda)$ 52

4.1. $Fp = \frac{\sum_{i=6}^{7} W_i P_i}{DP_w}$ 76

4.2. $Fp = \frac{W_7 P_7}{DP_w}$ 76

4.3. $Fn = \frac{\sum_{i=3}^{5} W_i P_i}{DP_w}$ 76

4.4. $Fn = \frac{\sum_{i=3}^{6} W_i P_i}{DP_w}$ 76

4.5. $Fn = \frac{(\sum_{i=3}^{5} W_i P_i)+W_1 P_{1,n}}{DP_w}$ 76

4.6. $Fm = \frac{\sum_{i=1}^{2} W_i P_i}{DP_w}$ 76

4.7. $Fm = \frac{(\sum_{i=1}^{2} W_i P_i)- W_1 P_{1,n}}{DP_w}$ 76

4.8. $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i^D - X_i^C\right)^2}$ 81

4.9. $rRMSE = 100\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i^D - X_i^C}{0.5(X_i^D - X_i^C)}\right)^2}\right)$ 81

4.10. $Bias = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^D - X_i^C\right)$ 81

4.11. $rBias = 100\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i^D - X_i^C}{0.5(X_i^D - X_i^C)}\right)\right)$ 81

4.12. $Fp = \frac{0.5W_6P_6 + W_7P_7}{DP_w}$ 90

4.13. $Fn = \frac{0.5W_1P_1 + 0.75W_2P_2 + \sum_{i=3}^{5} W_iP_i + 0.5W_6P_6}{DP_w}$ 90

4.14. $Fm = \frac{0.5W_1P_1 + 0.25W_2P_2}{DP_w}$ 90

5.1. $Diat_{\mathrm{DPA}} = \frac{1.41\mathrm{Fuco}}{\sum DP_w}$ 101

5.2. $DiatChl_{DPA} = Diat_{DPA} * (Chl\ a)$ 101

5.3. $DiatCarb_{DPA}\_mean = 41 * DiatChl_{DPA}$ 102

5.4. $DiatCarb_{DPA}\_max = 107 * DiatChl_{DPA}$ 102

5.5. $DiatCarb_{DPA}\_min = 15 * DiatChl_{DPA}$ 102

5.6. $DiatCarb_{DPA}\_range = DiatCarb_{DPA\_}\max - DiatCarb_{DPA}\_min$ 102

5.7. $\Delta Pigment = Pigment_{a_p} - Pigment_{global}$ 103

5.8. $DiatCarb_{Imagery}\_Unc = \sqrt{0.17^2 + 0.18^2 + 0.1^2 + 0.29^2}$ 108

5.9. $DiatCarb_{NN}\_Unc = \sqrt{0.39^2 + 0.52^2}$ 111

5.10. $DiatChl_{H11} = ([1.3272 + \exp(-3.9828x + 0.1953)]^{-1}) * \mathrm{Chl}\ a$ 114

5.11. $\Delta DiatCarb = DiatCarb_{H11} - DiatCarb_{NN}$ 114

LIST OF ABBREVIATIONS

$a^*(\lambda)$	Pigment-specific absorption
Allo	Alloxanthin
$a_p(\lambda)$	Particulate spectral absorption
$a_{tot}(\lambda)$ or $a(\lambda)$	Total spectral absorption
$a_{0.2\mu m}(\lambda)$	0.2µm-filtered spectral absorption
$a_{CDOM}(\lambda)$	Absorption by colored dissolved organic matter
$a_{gaus}(\lambda)$	Gaussian peak absorption
$a_{NAP}(\lambda)$	Non-algal particle spectral absorption
$a_\phi(\lambda)$	Phytoplankton absorption
$a_w(\lambda)$	Absorption by water
$b_b(\lambda)$	Spectral backscattering
$b_{bp}(\lambda)$	Particulate backscattering
$b_{bw}(\lambda)$	Backscattering by water
$b_p(\lambda)$	Particulate scattering
Buta	19'-butanoyloxyfucoxanthin
C:Chl	Carbon-to-chlorophyll ratio
Chl *a*	Chlorophyll *a*
Chl *b*	Chlorophyll *b*
Chl *c*	Chlorophyll *c*
$c_p(\lambda)$	Particulate attenuation
DMSP	Dimethylsulfoniopropionate
DP	Diagnostic pigments

DPA	Diagnostic Pigment Analysis
$E_d(\lambda)$	Downwelling irradiance
ESD	Equivalent spherical diameter
Fm	Microplankton fraction
Fn	Nanoplankton fraction
Fp	Picoplankton fraction
Fuco	Fucoxanthin
γ	Slope of particulate attenuation
Hexa	19'-hexanoyloxyfucoxanthin
HPLC	High Performance Liquid Chromatography
ICS	Influx Cell Sorter
IFCB	Imaging FlowCytobot
$K_d(\lambda)$	Downwelling attenuation coefficient
$K_{Lu}(\lambda)$	Upwelling attenuation coefficient
λ	Wavelength (nm)
$L_u(\lambda)$	Upwelling radiance
$L_w(\lambda,0^+)$	Water-leaving radiance
MODIS	Moderate Resolution Imaging Spectroradiometer
NAAMES	North Atlantic Aerosol and Marine Ecosystems Study
OCI	Ocean Color Instrument
PACE	Plankton Aerosol Cloud & Ocean Ecosystem
PCC	Phytoplankton community composition
Peri	Peridinin
PFTs	Phytoplankton functional types

PPC	Photoprotective carotenoids
PSC	Photosynthetic carotenoids
PSCs	Phytoplankton size classes
$R_{rs}(\lambda)$	Remote-sensing reflecting
SABOR	Ship-Aircraft Bio-Optical Research
TChl *a*	Total chlorophyll *a*
TChl *b*	Total chlorophyll *b*
μ_u	Average cosine for upwelling light
Zea	Zeaxanthin

CHAPTER 1

INTRODUCTION

1.1 Background

Phytoplankton are primary producers found ubiquitously throughout the sunlit surface waters of the world ocean and are the base of the marine food web. These microscopic, single-celled organisms produce oxygen as a byproduct of photosynthesis, and as a result phytoplankton contribute significantly to the oxygenation of Earth's atmosphere. Although microscopic, as a result of their immense numbers and high turnover rate, phytoplankton contribute approximately half of the Earth's net primary production (Field et al. 1998; Behrenfeld et al. 2001). Phytoplankton impact biogeochemical cycling of elements required for their growth, including nitrogen, phosphorus, and silica. Phytoplankton also play a role in the global carbon cycle via uptake of inorganic carbon in the form of carbon dioxide (CO_2) that enters the ocean via equilibration with the atmosphere, and carbon export that results from downward flux of dead and aggregated phytoplankton cells or fecal pellets of the organisms that prey on phytoplankton (Suess 1980; Alldredge and Silver 1988; Michaels and Silver 1988; Turner 2015). Other important phytoplankton functions include their influence on climate and atmospheric properties via production of dimethylsulfoniopropionate (DMSP) (Bates et al. 1987) and other volatile organic compounds, and their role in water and food (e.g., shellfish) quality resulting from harmful algal blooms.

Beyond bulk phytoplankton biomass is the importance of the underlying phytoplankton community composition (PCC) and its dynamics in both time and space. Phytoplankton are extremely diverse, both taxonomically and functionally, with highly variable sizes, morphologies, and mechanisms for growth and nutrient use. Knowledge of PCC distributions is necessary for addressing ocean ecosystem questions relevant to the size and composition of particles in the ocean, which in turn impact biogeochemical nutrient cycling, carbon export, and organisms in higher trophic levels that rely either directly or

indirectly on phytoplankton for nutrition. Phytoplankton community size structure is responsible for the variability in downward particle mass flux and carbon export, with increased flux in regions dominated by larger phytoplankton, and small eukaryotic plankton contributing significantly to carbon export in the oligotrophic ocean (Boyd and Newton 1995; Guidi et al. 2009; 2015). PCC is also linked to the variability in food quality of phytoplankton to higher tropic levels (i.e., food chain efficiency) (Dickman et al. 2008). Differences in algal food quality, such as the amount of lipid content, for zooplankton that result from changes in PCC are shown to be more important than changes in macronutrient composition (Brett et al. 2000). Additionally, biogeochemical models that include information on PCC will benefit from improved knowledge of PCC distributions for model development and testing (Bopp 2005; Follows et al. 2007; Dutkiewicz et al. 2013).

1.2 Defining Phytoplankton Community Composition

Taxonomic groups, phytoplankton size distributions, or functional groups with similar roles in biogeochemical nutrient cycling have all been used to define PCC. Phytoplankton include both prokaryotic (bacteria) and eukaryotic photosynthetic organisms representing several phyla (Table 1.1), with all eukaryotic chloroplasts originating evolutionarily in cyanobacteria and subsequently acquired via endosymbiosis by phytoplankton found in the Plantae, Chromista, and Protozoa kingdoms (Delwiche 1999; Sanchez-Puerta and Delwiche 2008). Hundreds of genera and thousands of species of phytoplankton have been described, and these numbers are continuously in flux as taxonomy is revised and new species are identified (Jeffrey et al. 1997; WoRMS Editorial Board 2020). In an effort to distill the many species of phytoplankton down to a number of groups more easily applied during ecosystem and modeling studies (e.g., 5-10 groups), Phytoplankton Functional Types (PFTs) are defined in the literature as groups of phytoplankton taxa that play similar functional roles in biogeochemical nutrient cycling, such as calcifiers (e.g., coccolithophores), silicifiers (e.g., diatoms), or nitrogen-fixing

cyanobacteria (Le Quere et al. 2005; IOCCG 2014; Nair et al. 2008). Phytoplankton cell size is also an important functional trait and discussed in more detail in the following paragraph.

The size distribution of cells, often described by phytoplankton size classes (PSCs), is sometimes used to describe PCC. Phytoplankton span a broad range of sizes and morphologies, from the smallest prokaryotic cyanobacteria that are < 1 μm to large dinoflagellates and chain-forming diatoms that are > 100 μm. Phytoplankton are often considered in broad size categories, originating with the work of Sieburth et al. (1978), who defined three logarithmically-scaled classes: picoplankton (< 2 μm), nanoplankton (2-20 μm), and microplankton (> 20 μm). Most major phytoplankton groups have taxa distributed across multiple size classes (Table 1.1). As many phytoplankton types are not spherical, especially in the micro- and large nanoplankton size range, the metric used to define cell size becomes more important. For example, a cell might fall within the nanoplankton size range if defined by its equivalent spherical diameter (ESD), and the microplankton size range if defined by its major axis length. For consistency throughout this thesis, "cell size" refers to ESD. ESD can be derived either from an estimated particle volume or from a sphere calculated from the cross-sectional area of a particle; here we use volume-based ESD. Additional morphological parameters of phytoplankton cells such as their eccentricity or surface area may be important to consider for processes such as nutrient uptake, light absorption, and predator-prey encounter rate. These properties can also be used as proxies or intrinsic properties to distinguish roles of different phytoplankton in the ecosystem.

Defining PFTs by proxies such as diagnostic pigments presents a challenge, as groups that have coherent biogeochemical roles may span a wide range of sizes and may not be associated with unique diagnostic pigments. For example, silicifiers in the ocean (predominantly diatoms, as well as silicoflagellates and radiolarians; the latter are protists that may have photosynthetic algal symbionts) fall within a wide size range (several to hundreds of μm), and do not contain accessory pigments to distinguish them from

other groups (fucoxanthin, which is commonly associated as a marker pigment for diatoms, is also found in pelagophytes and prymnesiophytes). However, phytoplankton grouped by functional role or size class can be more easily incorporated into biogeochemical models that also incorporate processes of nutrient cycling (e.g., Dutkiewicz et al. 2013).

Table 1.1. Taxonomic classification and size classes of common oceanic phytoplankton. Data from (Jeffery and Vesk 1997; Roy et al. 2011; www.marinespecies.org).

Kingdom	Division (Phylum)	Taxonomic Class	Common name and example common genera	Size Classes		
				Pico	Nano	Micro
Bacteria	Cyanobacteria	Cyanophyceae	*Synechococcus* *Trichodesmium* *Prochlorococcus*	✓		✓
Chromista	Heterokontophyta (a.k.a. Ochrophyta)	Bacillariophyceae	Diatoms		✓	✓
		Chrysophyceae	*Dinobryon*		✓	✓
		Pelagophyceae	Pelagophyte *Pelagomonas*	✓	✓	
		Dictyochophyceae	Silicoflagellate Dictyochales		✓	✓
	Haptophyta	Prymnesiophyceae	*Phaeocystis* Coccolithophores	✓	✓	✓
	Cryptophyta	Cryptophyceae	*Cryptomonas*		✓	
	Myzozoa (a.k.a. Dinophyta)	Dinophyceae	Dinoflagellates		✓	✓
Protozoa	Euglenozoa	Euglenoida	Euglenoids		✓	✓
Plantae	Chlorophyta	Pyramimonadophyceae	*Pyramimonas*		✓	
		Mamiellophyceae	*Micromonas*	✓	✓	✓
		Prasinophyceae	*Pterosperma*			

1.3 Measuring PCC on Global and Ocean Basin Scales

Methods for measuring PCC "directly", i.e., by analysis of water samples collected in the field, include microscopy, flow cytometry, and genomics techniques. Instruments deployed on shipboard flowing seawater systems can be used to collect in situ data at increased spatial and temporal resolution over discrete samples, and include light absorption and attenuation, fluorescence, particle size distribution, and imaging-in-flow cytometry. In addition to the knowledge gained through analysis of water samples, in situ measurements of PCC are also necessary to develop and evaluate algorithms that map PCC remotely. The use of proxies determined from remote sensing data to observe and study changes in PCC is a necessary tool that provides information on global and ocean basin scales otherwise unattainable. Total phytoplankton biomass is estimated on large scales using satellite-derived Chl *a* concentrations (e.g., Campbell and Aarup 1992; Doney et al. 2003; Antoine et al. 2005), and phytoplankton biomass dynamics are estimated by biogeochemical modeling efforts (e.g. Dutkiewicz et al. 2001; Behrenfeld et al. 2013). To enable the study of ocean ecosystems on large spatial scales, improved methods to detect PCC by proxy (e.g., via optics and by satellite), and quantify associated uncertainties, are needed. Scaling up from in situ measurements to ocean basin scale observations (Fig. 1.1) through the use of bio-optical or other modeling approaches requires careful consideration of both in situ measurement accuracies and uncertainties, as well as challenges of defining criteria for coincidence of different measurement types and comparisons of PCC defined in different units (e.g., pigments, biovolume, carbon, or cell counts).

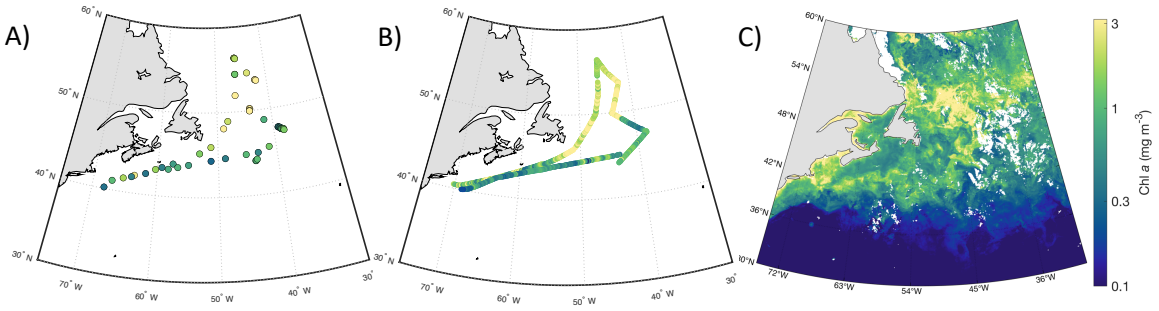

Figure 1.1. Example spatial resolution of different data types. Maps show A) discrete water samples, B) automated measurements made on water pumped through a shipboard flow-through system, and C) satellite estimates of chlorophyll *a* (Chl *a*) from ocean color radiometry. All three maps have the same color scale for Chl *a* concentration in mg m^{-3} as shown in the colorbar at the right. Panels A) and B) are in situ data collected during the NAAMES 02 cruise in May, 2016 with Chl *a* determined by quantitative pigment analysis and the magnitude of the red peak of particulate absorption, respectively. Panel C) is the MODIS Aqua L3 Chl *a* product from the NASA OC.DAAC for the month of May 2016.

With satellite data providing near-daily global coverage, numerous studies have worked towards identifying phytoplankton functional types or taxonomic groups using remote-sensing reflectance ($R_{rs}(\lambda)$) spectra measured from space. The variety of published approaches includes analysis of reflectance spectral anomalies and band-ratios (Alvain et al. 2005; 2008; Ben Mustapha et al. 2013; Sathyendranath et al. 2004; Kramer et al. 2018), retrievals of absorption spectra representing different phytoplankton groups using Differential Optical Absorption Spectroscopy (Bracher et al. 2009; Sadeghi et al. 2012), and spectral inversion models (Werdell et al. 2014; Westberry and Siegel 2005). Hirata et al. (2011) developed empirical relationships between Chl *a* and PCC both for size classes and taxonomic groups. Other studies have estimated PSCs and particle size distributions from space in an effort to map the contribution of different PSCs to total Chl *a*, using empirical relationships between

Chl *a* and in situ estimated PSCs (Uitz et al. 2006), and spectral features of $R_{rs}(\lambda)$ and associated derived products (Devred et al. 2011; Brewin et al. 2016; Li et al. 2013; Kostadinov et al. 2009).

Models developed to estimate phytoplankton groups from optical proxies (i.e., satellite ocean color data) are inherently limited by uncertainties in both the remote-sensing and in situ data used during algorithm development and evaluation. Currently, phytoplankton pigment concentrations measured using High Performance Liquid Chromatography (HPLC) analysis performed on discrete water samples are the most commonly used data type for definition of phytoplankton groups, usually through the assignment of diagnostic pigments. Other approaches to measure PCC include cell counts from microscopy (Werdell et al. 2014; Westberry and Siegel 2005) or Continuous Plankton Recorder (CPR) data (Raitsos et al. 2008). Information on phytoplankton accessory pigments including their absolute concentrations and quantities relative to Chl *a* have great utility in describing phytoplankton communities when full suites of HPLC-determined pigments are considered (e.g., Mackey et al. 1996; Swan et al. 2016; Kramer and Siegel 2019). A few individual pigments, namely alloxanthin, peridinin, and divinyl forms of chlorophyll are diagnostic for cryptophytes, dinoflagellates, and Prochlorophytes, respectively. However, many other pigments are not diagnostic but are present in multiple major phytoplankton groups (Table 1.2). Existing methods that rely on diagnostic pigments to define PCC – both size classes and taxonomic groups – are affected both by poorly constrained relationships between pigments and groups or PSCs, as well as inherent limitations in defining PCC in units of either relative or absolute pigment concentrations.

Most remote-sensing data used in space-based PCC algorithm development has been multispectral in nature, reflecting the currently available ocean color satellite instrument specifications (exceptions are the use of SCIAMACHY data in the PhytoDOAS routine – see Bracher et al. 2009; Sadeghi et al. 2012; Losa et al. 2017). However, upcoming hyperspectral (≤ 5 nm resolution) satellite ocean color

instruments such as the NASA PACE and German EnMAP missions will enable development of new PCC algorithms that make use of the added information available from hyperspectral data. Preceding the launch of hyperspectral satellite instruments, studies have made use of in situ hyperspectral measurements; for further discussion and literature review of studies using hyperspectral optical measurements to detect phytoplankton groups and pigments, see Chapter 2.

Table 1.2. Occurrence of pigments within phytoplankton groups. Darker box shading indicates the presence of a major pigment, while light shading indicates either minor, or sometimes present pigments (Jeffery and Vesk 1997; Roy et al. 2011). Note that biliproteins are not detectable using standard HPLC analysis. Chl = chlorophyll; Buta = 19'-butanoyloxyfucoxanthin; Hexa = 19'-hexanoyloxyfucoxanthin.

Common or taxonomic name	Chl a	Divinyl Chl a	Chl b	Divinyl Chl b	Chl c1	Chl c2	Chl c3	Alloxanthin	19'-Buta	Diadinoxanthin	Diatoxanthin	Fucoxanthin	19'-Hexa	Lutein	Neoxanthin	Peridinin	Prasinoxanthin	Violaxanthin	Zeaxanthin	Biliproteins
Synechococcus Trichodesmium	■																		■	■
Prochlorococcus		■	■	■															■	■
Diatoms	■				■	■	■			■	■	■								
Chrysophyceae	■				■	■	■			■	■	■							■	
Pelagophyceae	■				■	■	■		■	■	■	■	■							
Silicoflagellate	■				■	■	■		■	■	■	■								
Prymnesiophyceae	■				■	■	■		■	■	■	■	■							
Cryptophyceae	■					■		■												■
Dinoflagellates	■		■		■	■	■	■		■	■	■	■			■			■	
Euglenophytes	■		■							■	■				■				■	
Chlorophytes	■		■											■	■		■	■	■	
Prasinophytes	■		■											■	■		■	■	■	

1.4 Thesis Data and Objectives

Instrumentation and techniques to measure PCC, with an emphasis on optical, imaging, and cytometry methods, have advanced rapidly in recent years (Lombard et al. 2019). These measurements provide both increased phytoplankton taxonomic information (for example from cell imagery that allows for identification of cell types), and increased data coverage (mainly the result of high-temporal resolution shipboard and autonomous vehicle measurements). One recent investigation, the North Atlantic Aerosol and Marine Ecosystems Study (NAAMES), occurred over four seasons in 2015-2018 in the western North Atlantic (Behrenfeld et al. 2019). NAAMES used a multi-disciplinary approach to investigate the annual cycle of phytoplankton communities in a region of the world – the North Atlantic Ocean – known to have some of the most productive annual phytoplankton blooms in the world, and estimated to account for approximately one fifth of net global CO_2 uptake (Sabine et al. 2004). From the four NAAMES cruises, several data types were used in the research presented in Chapters 4 and 5. These data include "continuous" measurements made by instruments deployed in-line on the shipboard flowing seawater system: spectral particulate absorption and attenuation, and imaging-in-flow cytometry data for quantification of phytoplankton cell size and types in surface water. Measurements performed on discrete water samples and used in the present research include conventional flow cytometry analyses, and phytoplankton pigment concentrations from HPLC analysis. The analysis of Chapter 3 uses data collected from five independent investigations distributed across the globe. These global data do not contain the same extensive PCC information available from the NAAMES expedition, but have the advantage of broader global coverage.

Quantitative plankton imagery, such as is collected with the Imaging FlowCytobot (IFCB) instrument, provides novel opportunities to assess phytoplankton populations on time and space scales not previously possible. The majority of studies using IFCB data have been in coastal environments (e.g.,

Hunter-Cevera et al. 2016; Campbell et al. 2013), but open ocean studies (e.g., Laney and Sosik 2014) are likely to increase in the coming years as the use of automatic cell imaging methods increases (Lombard et al. 2019). During the NAAMES expedition, automated cell imagery was categorized using a deep learning network approach and the resulting data show the high spatial variability of several major phytoplankton groups in this productive region of the ocean (Fig. 1.2). The spatial variability, or "patchiness" of PCC in surface ocean waters is important to consider for in situ sampling campaigns that may have unknowingly biased results if the waters sampled do not well represent the larger region. Space-based methods to detect PCC also have the potential to shed light on the patchiness of phytoplankton populations.

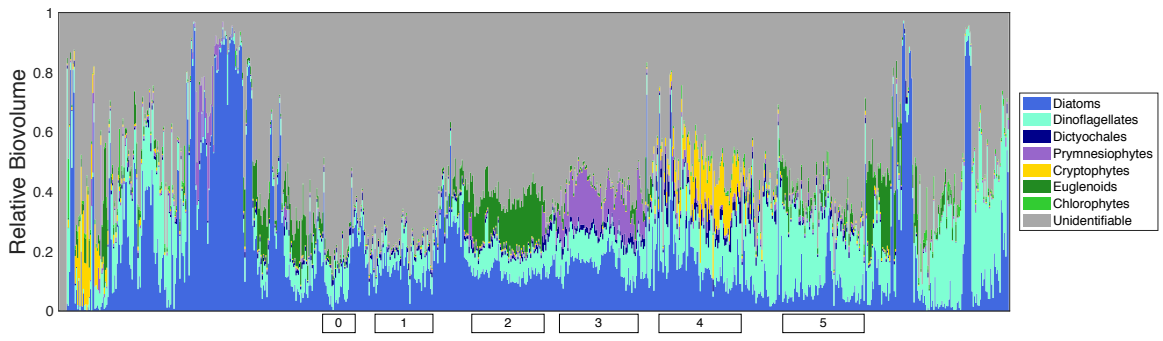

Figure 1.2. Relative biovolume of major identifiable phytoplankton groups. Data are from throughout the entire NAAMES 02 cruise (May 11 – June 5, 2016) in the western North Atlantic. The x-axis shows samples through time during ship transit, and the boxes show the number and duration of sampling station occupations. Biovolumes are calculated from cell imagery collected with the IFCB, which automatically sampled from surface (~ 5 m) waters. Values are calculated relative to the biovolume of all photosynthetic particles between 6 – 150 µm; note that this excludes small nanoplankton and all picoplankton from total cell biovolume.

Phytoplankton accessory pigments are a common theme throughout this thesis. Chapter 3 describes an approach for deriving phytoplankton accessory pigment concentrations from hyperspectral in situ measurements of $R_{rs}(\lambda)$, which has implications for upcoming hyperspectral ocean color satellite sensors. The main objective of this work is to demonstrate the use of hyperspectral $R_{rs}(\lambda)$ measurements for estimating accessory pigment concentrations, and the value added of these estimates over Chl *a*-based estimates. Based on research conducted for my Master's thesis (Chase et al. 2013) and by others, I hypothesize that as a result of the variable spectral absorption by different phytoplankton accessory pigments, the concentrations of pigments can be derived via spectral inversion and Gaussian decomposition, and that ratios of these accessory pigments to Chl *a* will show patterns in relative pigment concentrations that are independent of patterns that can be derived from pigment covariation relationships.

Chapters 4 and 5 evaluate previously published algorithms that use diagnostic pigments to define PCC and present updated and novel approaches to such algorithms. Chapter 4 applies a novel cytometry dataset to evaluate a commonly applied method to define PSCs from HPLC pigments, with the objective of quantifying uncertainties in the method and thus enabling more effective interpretation of its application. I hypothesize that the presence of dinoflagellates in the nanoplankton size class drives the overestimation of microplankton by the pigment-based approach, which traditionally attributes all peridinin – the major accessory pigment found in dinoflagellates – to the microplankton size class. More broadly, I hypothesize that there is high variability in the relationships between pigments and size class fractions, warranting careful consideration of the interpretation pigment-based PSCs. The information provided by combined conventional and imaging-in-flow cytometry will allow both quantification of uncertainties in the pigment-based PSCs, as well as necessary information for adjusting the pigment-based size class equations to reduce biases in results.

The main objective of Chapter 5 is to develop an algorithm to estimate diatom carbon from space-based measurements. An important component of this work is also to evaluate current published methods that estimate diatoms from pigments. I hypothesize that there is high variability in diatom carbon concentration as a function of Chl *a*, and that this variability can be modeled by incorporating both ocean color and environmental parameters into a neural network model. These hypotheses are addressable using the extensive NAAMES dataset of diatom cell imagery from across the western North Atlantic Ocean.

This thesis advocates both for the use of methods to estimate pigments from optical measurements, as well as careful consideration of pigment use to define PCC. These two themes are not in contradiction; the information provided by accessory pigments measured by HPLC or estimated from optics undoubtedly provides valuable information on PCC, but the use of pigments as the sole approach to defining PCC requires careful consideration. Additionally, and importantly, many pigments and phytoplankton groups in the ocean covary with Chl *a* concentration (Trees et al. 2000; Pan et al. 2010; Kramer and Siegel 2019; data of this thesis), and as a result it is necessary to quantify the value added an algorithm provides over what can be determined solely from relationships between phytoplankton groups or pigments, and Chl *a*. Chapters 3 and 5 demonstrate that there is a need to understand when and why phytoplankton pigments and communities deviate from expected relationships with Chl *a*. Overall, the work presented in the following chapters aims to describe the ways in which I have used novel data and approaches to develop algorithms that contribute to detecting PCC on broad scales in the ocean.

CHAPTER 2

PERSPECTIVES ON HYPERSPECTRAL OPTICAL MEASUREMENTS

2.1 Why Measure Optical Properties?

Solar radiation penetrating the ocean surface interacts with water molecules and the dissolved and particulate matter within the water. Photons entering the water can either be scattered (potentially multiple times), absorbed, or re-emitted in the case of fluorescence by photosynthetic organisms. The combined scattering and absorption results in the attenuation of photons; the term "light" will be used here to indicate photons between 400 and 700 nanometers (nm) in wavelength (λ); i.e., the visible wavelengths. Dissolved and particulate matter found in the ocean alter the light field differently depending on their composition and concentrations. An extensive body of work related to the field of optical oceanography has sought to characterize the light-attenuating properties of various constituents in the water (as well as water itself), and to describe and quantify the radiative transfer processes that link the attenuating constituents to the light field (e.g., Kirk 1994; Mobley 1994).

Measuring the light field within or above the water (e.g., by satellites) can provide greatly increased spatial and temporal information compared to quantifying all seawater constituents directly (e.g., via filtration of water samples). For this reason, the absorbed, scattered, attenuated, fluoresced, or reflected light – i.e., the optical properties – are routinely measured. Optical measurements can be made at either multi- or hyperspectral resolution. For the purposes of this chapter and following the literature (see Werdell et al. 2018 and references therein), hyperspectral measurements are defined by a spectral resolution of ≤ 5 nm across the visible wavelengths, and multispectral measurements by spectral resolution lower than that. Over the past two decades, a body of work within optical oceanography has addressed the potential benefits as well as some of the limitations specific to using hyperspectral data to estimate the relative and/or absolute quantities of light-attenuating materials in

the ocean. A selected review of those studies that address this challenge as it relates to detecting phytoplankton community composition (PCC) is presented here.

2.2 Why Hyperspectral?

Hyperspectral data provide, by definition, increased information over multispectral data, as an increased number of measurements are being made over the same spectral range. Several studies have addressed the topic of optimal spectral resolution and/or band placement for radiometric measurements. Wolanin et al. (2016) showed that band placement requirements depend on both the method as well as the target phytoplankton groups to be retrieved, and they suggest that hyperspectral (vs. any lower spectral resolution) observations are ideal. Although previous studies have advocated for band placement at spectral resolutions lower than 5 nm (Lee et al. 2007; Isada et al. 2015; Lee and Carder 2002), the results of Vandermeulen et al. (2017), who used a database of in situ hyperspectral reflectance measurements from a wide range of water types, showed that spectral resolution of 5 nm is optimal to separate differently absorbing phytoplankton groups while also accounting for measurement uncertainties. Hyperspectral phytoplankton absorption and remote-sensing reflectance ($R_{rs}(\lambda)$) spectra are more effective than multispectral data during derivative and clustering analyses for pigment assemblage discrimination and size-based, phytoplankton community composition assessment (Torrecilla et al. 2011; Uitz et al. 2015; Roelke et al. 1999). This finding, discussed in greater detail below, reflects the relative similarity of spectral absorption of different phytoplankton pigments and groups (e.g., Mao et al. 2010), which necessitates the use of optical information at high spectral resolution to discern subtle differences in spectral absorption and reflectance that are the result of differently absorbing phytoplankton pigments.

2.3 Deriving Information from Hyperspectral Measurements

Two general approaches can be taken when deriving information on PCC from optical measurements: 1) "direct" use of optical spectra, or 2) inversion methods to determine the light attenuation, absorption, or scattering of constituents in the water that contribute to changes in spectral shape. Various constituents in seawater act to attenuate light, including colored dissolved organic matter (CDOM; sometimes referred to in the literature as "gelbstoff" i.e., "yellow substance"), detrital material (also referred to an "non-algal particles" (NAP)), phytoplankton, and water molecules themselves (via absorption and Raman scattering). Phytoplankton, the main concern of this chapter, and its absorption ($a_\phi(\lambda)$), influences total absorption ($a(\lambda)$), where $a(\lambda)$ represents all absorbing constituents additively by:

$$a(\lambda) = a_\varphi(\lambda) + a_{CDOM}(\lambda) + a_{NAP}(\lambda) + a_w(\lambda) \quad (2.1)$$

and where $a_{CDOM}(\lambda)$, $a_{NAP}(\lambda)$, and $a_w(\lambda)$ are absorption by CDOM, NAP, and water, respectively. Variable absorption by different phytoplankton pigments in turn influences the shape of $R_{rs}(\lambda)$ spectra (Morel and Prieur 1977), and follows an inverse relationship between absorption and reflectance:

$$R_{rs}(\lambda) = \frac{f}{Q}\frac{b_b(\lambda)}{a(\lambda) + b_b(\lambda)}, \quad (2.2)$$

where f and Q are parameters related to measurement view angle, atmospheric conditions, and optical properties of seawater (Morel and Gentili 1996) and $b_b(\lambda)$ is spectral backscattering. These relationships between $a_\phi(\lambda)$, $a(\lambda)$, and $R_{rs}(\lambda)$ are the reason that optical measurements can be used to study phytoplankton absorption and from that, PCC, following the assumption that different $a_\phi(\lambda)$ indicates different phytoplankton types. Phytoplankton pigments are the physical link between optical measurements and phytoplankton cells. The spectra of $a_\phi(\lambda)$ are different among phytoplankton taxonomic groups as a result of the presence of various accessory pigments, which absorb more or less

light at different wavelengths (Fig. 2.1; Bidigare et al. 1990; Bricaud et al. 2004; Clementson and Wojtasiewicz 2019).

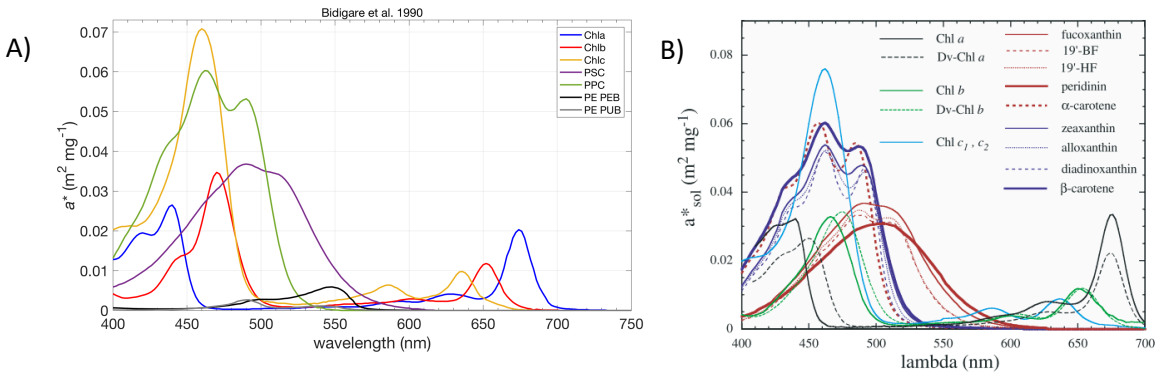

Figure 2.1. Pigment-specific absorption spectra ($a^*(\lambda)$) across the visible spectrum. A) Data from Table 6 in Bidigare et al. (1990) and represent measurements of pure pigment standards that have been wavelength-shifted to in-vivo maxima. B) Fig 1. from Bricaud et al. (2004); spectra are estimated from absorption by pigments in solvent, scaling the values to coefficients in Goericke and Repeta (1993), and wavelength-shifting to in-vivo maxima as in Bidigare et al. (1990). Chl = chlorophyll; PSC = photosynthetic carotenoids; PPC = photoprotective carotenoids; PE = phycoerythrin.

2.3.1 Direct Use of Optical Spectra

The direct use of hyperspectral optical measurements includes those methods in which spectra and/or their derivatives are linked via empirical relations to PCC metrics. Derivative analysis enhances the features in spectra that are due to absorption by different phytoplankton pigments; see Millie et al. (1997) for an early study demonstrating the use of derivative analysis to detect a dinoflagellate in coastal Gulf of Mexico waters. The magnitude of derivative absorption at wavelengths of minimum or maximum values can be compared to phytoplankton pigments or groups by correlation or similarity index (SI). A combined derivative and SI approach was used by Isada et al. (2015) to show significant

correlations between the second derivative of normalized $a_\phi(\lambda)$ spectra and diatom contribution to chlorophyll *a* (Chl *a*) determined by CHEMTAX pigment analysis, and Kirkpatrick et al. (2000) used an SI approach to link the presence of dinoflagellates with $a_p(\lambda)$. In a comparison of multispectral and hyperspectral data, Lubac et al. (2008) found that hyperspectral $R_{rs}(\lambda)$ could effectively distinguish *Phaeocystis* blooms from diatom blooms in coastal waters by analysis of the position of minimum and maximum peaks in the second derivative of $R_{rs}(\lambda)$. Importantly, the authors tested and found that Chl *a* concentrations were not a significant factor determining their results. They also showed that CDOM concentrations did impact their results, and hence the need to estimate CDOM concentrations in coastal waters where concentrations are often non-negligible (Lubac et al. 2008). Xi et al. (2015) used a clustering and fourth-derivative approach to determine six phytoplankton taxonomic groups, with input data of $a_\phi(\lambda)$, $a_w(\lambda)$, and $R_{rs}(\lambda)$. Their results were improved when using $a_\phi(\lambda)$ compared to $R_{rs}(\lambda)$, and there were also improvements if wavelengths of strong water absorption were excluded and only samples with Chl $a > 1$ mg m^{-3} were considered. Based on their findings, the authors suggest direct use of $R_{rs}(\lambda)$ via combined derivative and clustering analysis rather than a two-step process that combines inversion to $a_\phi(\lambda)$ with subsequent derivative analysis of $a_\phi(\lambda)$ spectra. A related study by Xi et al. (2017) simulated $R_{rs}(\lambda)$ using Hydrolight radiative transfer software for five phytoplankton groups, and then applied a combined derivative and SI analysis to compare results to $a_\phi(\lambda)$ from cultured phytoplankton. The method was applied successfully to inland waters where Chl *a* concentrations were high and there was dominance of one phytoplankton group.

In addition to phytoplankton taxonomic groups, derivative and clustering analyses using $a_\phi(\lambda)$, $a(\lambda)$, and/or $R_{rs}(\lambda)$ have been used to estimate phytoplankton accessory pigments and pigment groups; many studies estimate pigments without explicitly linking results to the presence of different phytoplankton groups. One recent study found unique statistical relationships between first and second derivatives of $a_p(\lambda)$ spectra and pigment clusters in coastal waters off of California (Catlett and Siegel 2018).

Approaches aimed at deriving pigments in open-ocean waters include Torrecilla et al. (2011), who found that unsupervised hierarchical cluster analysis between pigment assemblages from HPLC analysis are linked with the second derivative of $a_\phi(\lambda)$ from field measurements. However, to use their method at another location or time, in situ data are needed to link clusters of pigments with clusters of optical data. See also Shaju et al. (2015) for an approach using 4th derivative analysis of $a_p(\lambda)$ to estimate pigments. Bracher et al. (2015) used an empirical orthogonal function (EOF) analysis to derive linear regression models between pigments and $R_{rs}(\lambda)$ spectra and found predictive capabilities for Chl *a* and several accessory pigments in the Atlantic Ocean. Uitz et al. (2015) also used open ocean $a_\phi(\lambda)$ and $R_{rs}(\lambda)$ data to show that combined clustering and spectral derivative as well as EOF analyses can be used to classify waters into bio-optical and biogeochemical categories. Relative concentrations of phycoerythrin (PE), a biliprotein found in cyanobacteria and cryptophytes, was estimated using a Generalized Linear Model with PE concentrations as the model target and EOF loadings from hyperspectral upwelling radiance ($L_u(\lambda)$) measurements made in water as the model input by Taylor et al. (2013). Derivative analysis can also be combined with inversion methods, as shown by Craig et al. (2006), who first inverted $R_{rs}(\lambda)$ to $a_\phi(\lambda)$ and then compared derivative analysis of $a_\phi(\lambda)$ to the absorption spectrum of the toxic dinoflagellate *Karenia brevis*, using a similarity index to show that $R_{rs}(\lambda)$ can indicate the bloom strength of *K. brevis*. Machine learning or neural network approaches to estimate pigments have received some attention (Bricaud et al. 2007; El Hourany et al. 2019; Deng et al. 2019) and more are likely in the future given the rapid adoption of artificial intelligence in all areas of science and technology.

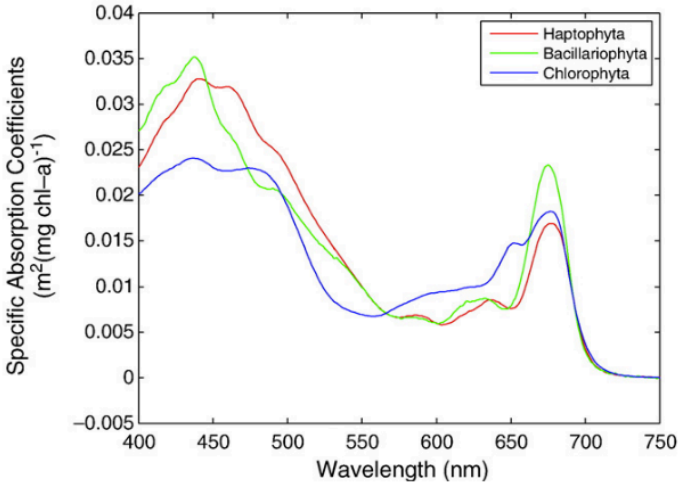

Figure 2.2. Specific absorption spectra of three representative phytoplankton phyla. (1) *Isochrysis* sp. representing Haptophyta (2) *Chaetoceros calcitrans* representing Bacillariophyta and (3) *Tetraselmis* sp. representing Chlorophyta (Fig. 4 from Mao et al. 2010).

2.3.2 Methods of Spectral Inversion

As outlined above, the spectral shapes of $R_{rs}(\lambda)$ and $a(\lambda)$ are a function of the absorbing and scattering components in the water (and the atmosphere, in the case of above-water measurements of $R_{rs}(\lambda)$). Methods to invert optical measurements to their component spectra, or eigenvectors, require *a priori* knowledge or assumptions about the shape of the component eigenvectors. Inversion methods can then be used to solve for the magnitude of eigenvectors that best recreate a measured spectrum. Inversion from $R_{rs}(\lambda)$ to bulk absorbing components including $a_\phi(\lambda)$ using in situ measurements has been explored extensively (e.g., Hoge and Lyon 1996; Roesler and Perry 1995; Werdell et al. 2013). Werdell et al. (2018) provide a review of methods to discern component absorbing and scattering spectra by inversion (and other methods). The pigment-specific absorption ($a_\phi^*(\lambda)$; Yentsch and Phinney 1989; Sosik and Mitchell 1995), which can be considered an eigenvector of the bulk phytoplankton population, is highly variable

as a result of phytoplankton species composition, cell size, and growth conditions (Bricaud et al. 2004; Ferreira et al. 2013; Poulin et al. 2018). Extending inversion beyond bulk phytoplankton absorption to eigenvectors representing phytoplankton groups presents a challenge because the problem is likely to be ill-posed; that is, there are likely to be multiple solutions that satisfy the inversion equation resulting from different combinations of eigenvectors. Another consideration is the choice of eigenvectors – for example, if a phytoplankton group is assumed to be present during inversion but in fact is not, the eigenvector representing that group may be incorrectly assigned a positive magnitude. Additionally, growth conditions (e.g., light and nutrients) can change the spectral shape of absorption for the same phytoplankton group (Morel and Bricaud 1981; Ciotti et al. 2002; Organelli et al. 2017), further complicating the use of eigenvectors.

Studies applying inversion from hyperspectral $R_{rs}(\lambda)$ to phytoplankton groups (e.g., Roesler et al. 2004) are limited, whereas inversion from either $R_{rs}(\lambda)$ or $a_p(\lambda)$ to phytoplankton pigments is a more common approach. Estimating pigment concentrations from optical measurements is seemingly more straightforward than estimating phytoplankton groups, as pigments themselves directly influence the spectral absorption features. Conversely, determination of absolute or relative pigment concentrations from spectra still leaves the challenge of linking the pigment information to different phytoplankton groups, which is often non-trivial depending on which pigments are estimated, and which groups are of interest.

To estimate pigments from inversion of optical spectra, methods defining individual pigments by a single eigenvector (e.g., Moisan et al. 2011), or by a series of Gaussian functions have both shown success. Both Liu et al. (2019) and Ye et al. (2019) compared the eigenvector and Gaussian methods (Fig. 2.3) to show that both are effective at estimating chlorophyll or grouped carotenoids, but eigenvectors are more successful for separating some individual carotenoid pigments. In the Gaussian decomposition

approach, Gaussian peak assignment can be determined either from known local maxima of absorption pigments from laboratory measurements (Hoepffner and Sathyendranath 1991; Hoepffner and Sathyendranath 1993; Lohrenz et al. 2003; Chase et al. 2013), or by using derivative analysis to identify the location of peaks and "shoulders" in spectra (e.g., Aguirre-Gomez et al. 2001). Gaussian decomposition methods can be used to predict chlorophylls *a*, *b*, *c*, grouped carotenoid pigments, and PE in both open ocean waters and inland waters (Chase et al. 2013; 2017; Wang et al. 2016).

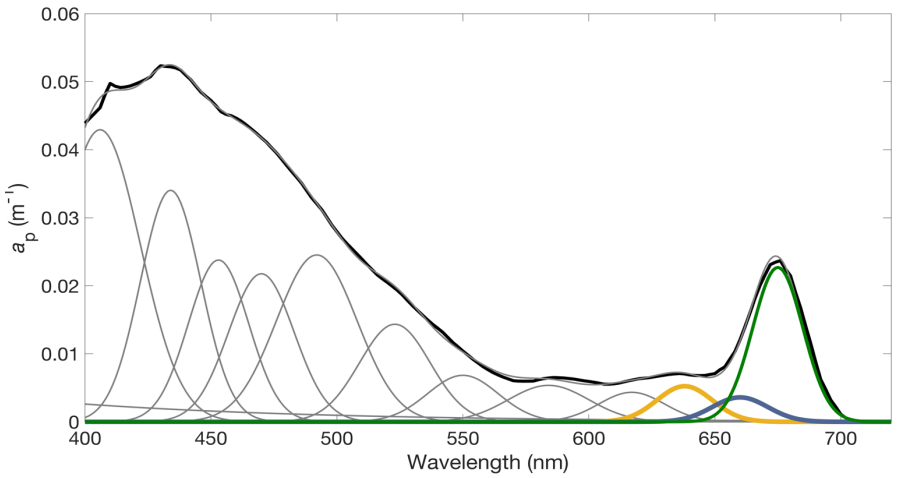

Figure 2.3. Example particulate absorption ($a_p(\lambda)$, m^{-1}) spectrum with component Gaussian functions. Gaussian magnitudes are derived following the spectral decomposition method of Chase et al. (2013). Curves representing absorption by Chl *a*, Chl *b*, and Chl *c* are shown by the green, dark blue, and yellow curves, respectively.

Table 2.1. Summary of previous studies estimating phytoplankton groups or pigments from hyperspectral optical measurements. Studies are divided into the two major categories of either direct use of spectra or inversion-based approaches. Details such as coastal vs. open ocean applications are omitted.

Approach	Input measurements	Result/product	Target/validation data	Reference
Direct use of optical measurements: Similarity Index, EOF, and/or clustering analysis	$a_\phi(\lambda)$ & 4th derivative of spectra	% contribution of *G. breve*	*G. breve* field and culture data	Millie et al. 1997
	2nd derivative of $a_\phi(\lambda)$	Diatom contribution to Chl *a*	CHEMTAX diatom Chl *a*	Isada et al. 2015
	$a_p(\lambda)$	Cell counts and Chl *a* fraction of *G. breve*	*G. breve* field and culture data	Kirkpatrick et al. 2000
	2nd derivative of $R_{rs}(\lambda)$	Detection of *Phaeocystis* blooms	Microscopic identification of phytoplankton	Lubac et al. 2008
	4th derivative of $a_\phi(\lambda)$ and $R_{rs}(\lambda)$	Differentiation of phytoplankton groups; cyanobacteria dominance in inland waters	Cultures, Hydrolight simulations, field $R_{rs}(\lambda)$ measurements	Xi et al. 2015; 2017
	Derivatives of $a_p(\lambda)$ or $a_\phi(\lambda)$	Pigment assemblages or concentrations	HPLC pigments or Chl *a* concentration from fluorescence	Catlett and Siegel 2018; Shaju et al. 2015; Torrecilla et al. 2011
	$R_{rs}(\lambda)$	Pigment concentrations	HPLC pigments	Bracher et al. 2015
	$a_\phi(\lambda)$ and $R_{rs}(\lambda)$, and derivatives	Bio-optical water categories	HPLC pigments	Uitz et al. 2015
	$L_u(\lambda)$	Relative phycoerythrin concentrations	PE concentration	Taylor et al. 2013
	$a_\phi(\lambda)$ and $R_{rs}(\lambda)$, and $a_\phi(\lambda)$ derivatives	*K. brevis* relative bloom strength	*K. brevis* absorption spectrum	Craig et al. 2006

Table 2.1 continued

Methods of spectral inversion: Spectral inversion and Gaussian decomposition	$a_p(\lambda)$ or $a_\phi(\lambda)$	Pigment concentrations or absorption	HPLC pigments	Aguirre-Gomez et al. 2001; Chase et al. 2013; Hoepffner and Sathyendranath 1991, 1993; Liu et al. 2019; Lohrenz et al. 2003; Ye et al. 2019
	$R_{rs}(\lambda)$	Contribution of phytoplankton groups to absorption	Microscopic cell counts	Roesler et al. 2004
	$R_{rs}(\lambda)$	Pigment concentrations	HPLC pigments	Chase et al. 2017; Wang et al. 2016

2.4 Making the Leap to Space-based Measurements

Errors and uncertainties are generally reduced when using $a_\phi(\lambda)$ or $a_p(\lambda)$ to estimate either phytoplankton groups or pigments compared to the use of $R_{rs}(\lambda)$ measurements, as there are more absorbing and scattering components to account for when using $R_{rs}(\lambda)$ (Eq. 2.1, 2.2). In addition, the addition of more absorbing constituents including pigments results in lower signal-to-noise (SNR) in the $R_{rs}(\lambda)$ signal, and the absorption by water in the red wavelengths also reduces SNR in that region of the spectrum. The advantage of measuring and using $R_{rs}(\lambda)$ for algorithm development is more direct application to satellite ocean color measurements, which are analogous to in situ $R_{rs}(\lambda)$ measurements following correction for atmospheric effects. Hyperspectral $R_{rs}(\lambda)$ can also be made using aircraft instruments, such as NASA's airborne visible/infrared imaging spectrometer (AVIRIS) instruments. The Hyperspectral Imager for the Coastal Ocean (HICO) instrument was installed on the International Space Station and provided hyperspectral ocean color measurements of coastal areas from 2009-2014 (Lucke et al. 2011). Data have demonstrated the use of hyperspectral $R_{rs}(\lambda)$ to observe a phytoplankton bloom in Monterey Bay, CA (Ryan et al. 2014), and to show that hyperspectral $R_{rs}(\lambda)$ can better estimate Chl *a*

compared to multispectral $R_{rs}(\lambda)$ in turbid waters and have the potential for distinguishing other pigments as well (Gitelson et al. 2011, Fig. 2.4). HICO-measured $R_{rs}(\lambda)$ have been used to map different phytoplankton species across a HICO image captured off the coast of China using a machine/transfer learning approach (Zhu et al. 2019).

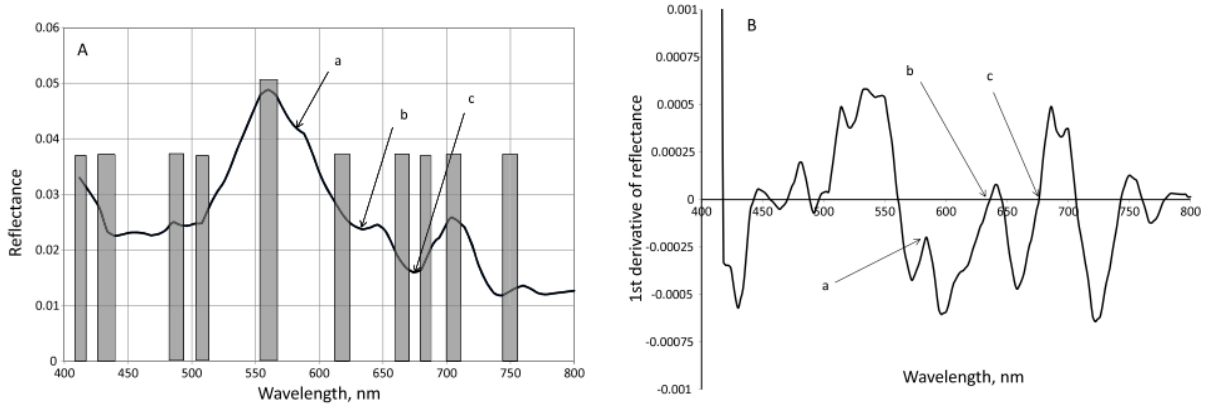

Figure 2.4. Median reflectance spectrum taken in the Azov Sea with marked locations of spectral bands of MERIS (A) and the first derivative of reflectance (B). Absorption bands of a-phycoerythrin; b-phycocyanin, and c-chlorophyll-a (Fig. 7 from Gitelson et al. 2011).

Uncertainties in $R_{rs}(\lambda)$ measurements arise when considering satellite-based measurements (vs. in situ or aircraft-based), as a result of the need to account for absorption and scattering taking place in the atmosphere. Propagating sensor noise in $R_{rs}(\lambda)$ measurements made from space not only requires information on sensor SNR, but also on the conditions during the which the measurements are made, including atmospheric properties, sensor view angle, and satellite altitude (Gillis et al. 2018). Propagating uncertainties in ocean color remote sensing products (e.g., Seegers et al. 2018; McKinna et al. 2019) is crucial to the development off PCC algorithms from space. Specific to hyperspectral measurements, calculating the value-added over multispectral is also crucial. One approach to this is to compare how products derived from hyperspectral compare to those which can be estimated from covariations with Chl *a* concentrations, as Chl *a* can be estimated from multispectral measurements.

Finally, consideration of the true number of degrees of freedom (DoF) in hyperspectral measurements is critical for algorithm development. Although hyperspectral measurements may have upwards of 60 wavelength values, the DoF are much lower than that, as a result of the relatively consistent shape of phytoplankton absorption in the ocean and the relatively small changes in spectral shape by differently absorbing pigments compared to measurement uncertainties. Hyperspectral particulate absorption spectra measured in situ have order ~ 5 DoF, and this number is not expected to be any higher for $R_{rs}(\lambda)$ measurements made from space (Cael et al. 2020). While seemingly limiting, this information can lead to improved hyperspectral PCC algorithms if it encourages the incorporation of ancillary information, such as water mass history or environmental parameters.

2.5 Take-home Points

1. Empirical or "direct" use of hyperspectral spectra and derivatives have significant relationships with phytoplankton pigments or groups (themselves often defined by pigment proxies) in many studies; however, evaluation to see if results are a significant improvement over what can be estimated from Chl *a* alone is necessary and often omitted.

2. Inverting hyperspectral spectra to retrieve component absorption spectra of different phytoplankton groups is generally ill-posed, unless prior knowledge of the groups present can be used to constrain the inversion.

3. Hyperspectral measurements are effective for quantifying phytoplankton accessory pigments; the number and type of pigments depends on measurement type and associated uncertainties.

4. Considerations of the DoF in hyperspectral measurements are critical when extracting information from spectra.

CHAPTER 3

ESTIMATION OF PHYTOPLANKTON ACCESSORY PIGMENTS FROM HYPERSPECTRAL REFLECTANCE SPECTRA: TOWARD A GLOBAL ALGORITHM

3.1. Introduction

The diversity of phytoplankton community composition in the ocean supports a wide range of ocean ecosystems. The ability to observe and monitor this diversity is necessary for understanding complex ocean processes, such as carbon export (e.g., Guidi et al. 2015; Mouw et al. 2016). Numerous methods are used to observe phytoplankton in the oceans, and all have benefits and limitations. Optical data, including spectral absorption and reflectance measurements, have the capability to be sampled at higher spatial and temporal resolution compared to analysis of discrete water samples. However, optical data also present the challenge of indirect observation of phytoplankton metrics such as biomass, composition, or pigments. Optical measurements made either in situ or remotely must be compared with coincidently measured data such as pigment concentrations or phytoplankton imagery to establish the utility of these data and the associated uncertainties. Remote-sensing reflectance ($R_{rs}(\lambda)$; λ represents wavelength) can be estimated both from in situ data collected using radiometers and remotely by ocean color (OC) sensors on board aircraft and satellites. Estimation of chlorophyll *a* (Chl *a*) concentration, the major phytoplankton pigment used during photosynthesis, is well established from $R_{rs}(\lambda)$ measurements using methods such as the band-ratio model (Gordon et al. 1983) and a three-band reflectance differencing method (Hu et al. 2012).

In addition to Chl *a*, phytoplankton groups contain different assemblages of accessory pigments for both photosynthetic and photoprotective purposes; these include chlorophylls *b* and *c*, carotenoids, and biliproteins. Determining the presence of different accessory pigments can be used to help characterize phytoplankton community composition (e.g., Mackey et al. 1996). Although different algal groups may contain the same pigments, some distinctions between them can be made based on certain pigments.

For example, chlorophyll *b* (Chl *b*) is found in green algae (Chlorophyceae and Prasinophyceae), whereas chlorophyll *c* (Chl *c*) is found in the golden-brown algae, which is a broad group including diatoms, dinoflagellates, silicoflagellates, and prymnesiophytes (Jeffrey & Vesk, 1997). Several xanthophyll pigments, a category of the carotenoids, are exclusive to distinct phytoplankton groups: alloxanthin is found in Cryptophyta; 19'-hexanoyloxyfucoxanthin is found in Prymnesiophyceae (including coccolithophorids); peridinin is found in dinoflagellates; and fucoxanthin is found in diatoms and golden-brown flagellates (which includes coccolithophorids and silicoflagellates) (Jeffrey & Vesk, 1997). The carotenoid pigments can also be grouped together as photoprotective (PPC) and photosynthetic (PSC) carotenoids; these two groups are made up of pigments that both have related roles in the cell and have similar light absorption spectra (e.g., Figure 1 in Bricaud et al. 2004). Previous efforts to move beyond the detection of Chl *a* and retrieve information about accessory pigments or different phytoplankton groups directly are described below, following a brief description of the theoretical relationship between pigment absorption and $R_{rs}(\lambda)$.

The total absorption ($a(\lambda)$) and backscattering ($b_b(\lambda)$) by all components in the water, including phytoplankton pigments, are determinants of $R_{rs}(\lambda)$ and their relationship is approximated (while ignoring inelastic scattering) by

$$R_{rs}(\lambda) = \frac{f}{Q}\frac{b_b(\lambda)}{a(\lambda) + b_b(\lambda)}. \quad (3.1)$$

The parameters *f* and *Q* relate measurements made at one angle to the nadir direction and are a function of sun angle, atmospheric conditions, and the optical properties of the water (Morel & Gentili, 1996). Total phytoplankton absorption ($a_\phi(\lambda)$) includes the absorption spectra of Chl *a* as well as all accessory pigments present. With multiple, varying phytoplankton pigments affecting $a(\lambda)$ and $a(\lambda)$

influencing $R_{rs}(\lambda)$, it has been the goal of several studies to estimate phytoplankton accessory pigments from $a(\lambda)$ or $R_{rs}(\lambda)$. Hyperspectral particulate absorption data ($a_p(\lambda)$), with information provided every few nanometers, have been used to estimate phytoplankton pigments from spectra measured in the laboratory (Hoepffner & Sathyendranath, 1991; 1993; Lohrenz et al. 2003; Moisan et al. 2011) and in situ (Chase et al. 2013). Extending a similar analysis to hyperspectral $R_{rs}(\lambda)$ data is more challenging and is complicated by the need to account for additional parameters that do not come into play with direct analysis of particulate or phytoplankton absorption spectra. These include inelastic scattering, particulate backscattering ($b_{bp}(\lambda)$) and absorption by dissolved matter and non-algal particles ($a_{CDOM}(\lambda)$ and $a_{NAP}(\lambda)$, respectively). In addition, there are uncertainties in the relationship between $R_{rs}(\lambda)$ and absorption coefficients (e.g., due to variability in f and Q in equation (3.1)).

Pigment-based clusters representing different phytoplankton assemblages were previously compared with clusters of absorption spectra, the second derivative of absorption spectra, and $R_{rs}(\lambda)$ data by Torrecilla et al. (2011). The authors found that both the absorption data and the second derivative of $R_{rs}(\lambda)$ compared well with pigment data, suggesting the utility of the optical measurements for describing phytoplankton pigments in the ocean. However, to use their method at another location or time, in situ data are needed to link clusters of pigments with clusters of optical data. Bracher et al. (2015) used an empirical orthogonal function (EOF) analysis to derive linear regression models between pigments and $R_{rs}(\lambda)$ spectra and found strong predictive capabilities for total Chl *a* (TChl *a*) and several accessory pigments in the Atlantic Ocean. A study by Pan et al. (2010) calculated ratios between $R_{rs}(\lambda)$ wavelengths from in situ data to develop algorithms for pigment prediction that were then applied to multispectral satellite data for the U.S. northeast coastal region. They used High Performance Liquid Chromatography (HPLC) pigment data for model development and validation and were able to estimate phytoplankton pigment concentrations from SeaWiFS and MODIS satellite $R_{rs}(\lambda)$ with mean absolute percent error for most pigments falling between 30% and 50%. Wang et al. (2016) estimated several

pigment absorption coefficients from inversion of in situ hyperspectral $R_{rs}(\lambda)$ data for inland lake waters containing high chlorophyll values and cyanobacterial bloom conditions (Chl *a* > 10 µg L^{-1}).

Another approach for the identification of phytoplankton groups in the ocean is to explore spectral anomalies and residuals that remain after removal of an average chlorophyll-based spectrum. This approach addresses the high covariation among phytoplankton pigments that has been observed on a global scale (Trees et al. 2000) and is used to extract information beyond the average relationship between Chl *a* and its covarying parameters, including accessory pigments. Brown et al. (2008) determined that $a_{NAP}(\lambda)$ and the magnitude of $b_b(\lambda)$ are responsible for the spread around the mean $R_{rs}(\lambda)$ – Chl *a* relationship, while Alvain et al. (2012) found that the Chl *a*-specific phytoplankton absorption ($a^*(\lambda)$), $a_{CDOM}(\lambda)$, and $b_b(\lambda)$ all influence reflectance anomalies. Ben Mustapha et al. (2013) built off previous work to improve the PHYSAT method (Alvain et al. 2005; 2008) to use multispectral satellite radiance anomalies to detect the dominance of several phytoplankton groups in the open ocean.

Many previous studies for detecting phytoplankton groups and pigments were conducted using multispectral reflectance data and/or were developed for a limited geographical region (e.g., Alvain et al. 2005, 2008; Ben Mustapha et al. 2013; Farikou et al. 2015; Raitsos et al. 2008; Werdell et al. 2014; also see Mouw et al. 2017 for a recent review). However, there is a need to understand both the added value and limitations of hyperspectral data and global algorithms, particularly given the increased spectral resolution of next-generation ocean color satellites (e.g., NASA's Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission and Germany's EnMAP mission). Algorithms that estimate accessory pigments over a broad global range of water types and without prior knowledge of the phytoplankton community composition may be preferred when global hyperspectral satellite $R_{rs}(\lambda)$ data become available within the next decade.

In this study, we explore the utility of hyperspectral $R_{rs}(\lambda)$ for estimating phytoplankton pigments using a global database of in situ $R_{rs}(\lambda)$ and HPLC measurements. We attempt to exploit the spectral differences in absorption by different phytoplankton accessory pigments, which in turn should influence the spectral shape of $R_{rs}(\lambda)$. As a comparison, we also examine the covariation between Chl *a* and accessory pigments and the capabilities of estimating accessory pigments using covariation relationships. Finally, we conduct a brief analysis of spectral residuals defined by deviations from the average global relationship between Chl *a* and $a_p(\lambda)$. These approaches help us understand the potential and the limitations for extracting information on phytoplankton pigments from hyperspectral $R_{rs}(\lambda)$ data on a broad global scale.

3.2. Data and Methods

3.2.1. Data Sets

Data sets from five different expeditions are used in our study (Table 3.1). For all data sets described below, HPLC pigment data were collected using discrete surface water samples (depth ≤ 5 m) that were filtered and preserved on board. When duplicate or triplicate HPLC samples were available the mean value is used. We studied four HPLC pigment groups (Table 3.2). $a_p(\lambda)$ data from an ac-s spectrophotometer (WET Labs, Inc., Philomath, OR, USA) deployed in a flow-through setup (Slade et al. 2010) are available for three of the expeditions (Tara Oceans, Tara Mediterranean, and SABOR). The ac-s data measured coincidently with $R_{rs}(\lambda)$ are used for calculating an attenuation correction when processing $R_{rs}(\lambda)$ spectra; see section 3.2.3 for ac-s and $R_{rs}(\lambda)$ processing details.

3.2.1.1. Tara Expeditions

Data from two extended expeditions on the R/V Tara are used: Tara Oceans (including Tara Oceans Polar Circle; global coverage from 2009–2013; Boss et al. 2014; Picheral et al. 2014) and Tara Mediterranean (Mediterranean Sea; June–September 2014). $R_{rs}(\lambda)$ spectra and HPLC pigment data collected within four hours of each other were used in the inversion analysis. There are 58 data points of coincident $R_{rs}(\lambda)$ and

HPLC data from varied water types around the globe from the Tara Expeditions (Figure 3.1a). The complete HPLC data sets from the Tara Expeditions are also used in our development of global pigment covariation relationships (n = 196; Figure 3.1b); these data are independent from the data used in the inversion analysis. Laboratory analysis of HPLC data was carried out at Laboratoire d'Oceanographie de Villefranche-sur-Mer (LOV) according to the method described by Ras et al. (2008), which is adapted from Van Heukelem and Thomas (2001). Finally, the global data set of ac-s spectral absorption measurements from the Tara Expeditions was used to calculate the shape and magnitude of average global particulate absorption spectra used in the residual analysis (Figure 3.1c; n = 96,929; see section 3.2.5).

3.2.1.2. SABOR, AE1319, and NH1418 Expeditions

Coincident HPLC and $R_{rs}(\lambda)$ data from three additional expeditions were used in the inversion analysis: SABOR (Gulf of Maine/North Atlantic/Mid-Atlantic coast; July–August 2014), AE1319 (North Atlantic and Labrador Sea; August–September 2013), and NH1418 (Equatorial Pacific; September–October 2014); see Figure 3.1a for data locations. $R_{rs}(\lambda)$ spectra and HPLC pigment data collected within 4 hours of each other were used in the inversion analysis. The Ocean Ecology Laboratory at NASA Goddard Space Flight Center performed HPLC analysis for the SABOR, AE1319, and NH1418 expeditions following methods in Van Heukelem and Thomas (2001) and further described in Hooker et al. (2009). The three expeditions combined contribute 39 matching HPLC and $R_{rs}(\lambda)$ data points (Table 3.1).

Table 3.1. Summary of datasets used in the present chapter

Expedition	Dates	Location(s)	Data type and number of points			
			$R_{rs}(\lambda)$ and HPLC for inversion analysis	HPLC global co-variation	$a_p(\lambda)$ for residual analysis	$a_p(\lambda)$ for $R_{rs}(\lambda)$ processing

Table 3.1. continued

Tara Oceans[a]	2009 – 2013	Global	29	196	89,421	25
Tara Mediterranean[b]	June – September 2014	Mediterranean Sea	29		7,508	29
SABOR[b]	July – August 2014	Gulf of Maine/North Atlantic/Mid-Atlantic coast	12			12
AE1319[b]	August – September 2013	North Atlantic and Labrador Sea	13			
NH1418[b]	September – October 2014	Equatorial Pacific	14			
Totals			**97**	**196**	**96,929**	**66**[c]

[a]Includes Tara Oceans Polar Circle Expedition; data available at:

https://doi.pangaea.de/10.1594/PANGAEA.836318 and

https://doi.pangaea.de/10.1594/PANGAEA.836319.

[b]Data available at https://seabass.gsfc.nasa.gov/ [*Werdell et al.* 2003].

[c]See section 3.2.2 for an explanation of $R_{rs}(\lambda)$ processing for the 31 data points without matching $a_p(\lambda)$ data.

Table 3.2 Pigment groups from HPLC analysis

Pigment group	Abbreviation	Sum
Total Chlorophyll *a*	TChl *a*	Monovinyl Chl *a* + divinyl Chl *a* + chlorophyllide *a* + Chl *a* allomers + Chl *a* epimers
Total Chlorophyll *b*	TChl *b*	Monovinyl Chl *b* + divinyl Chl *b* + Chl *b* epimers
Chlorophylls c_1 and c_2	Chl c_1+c_2	Chlorophyll c_1 + chlorophyll c_2
Photoprotective carotenoids	PPC	αβ-carotene + zeaxanthin + alloxanthin + diadinoxanthin

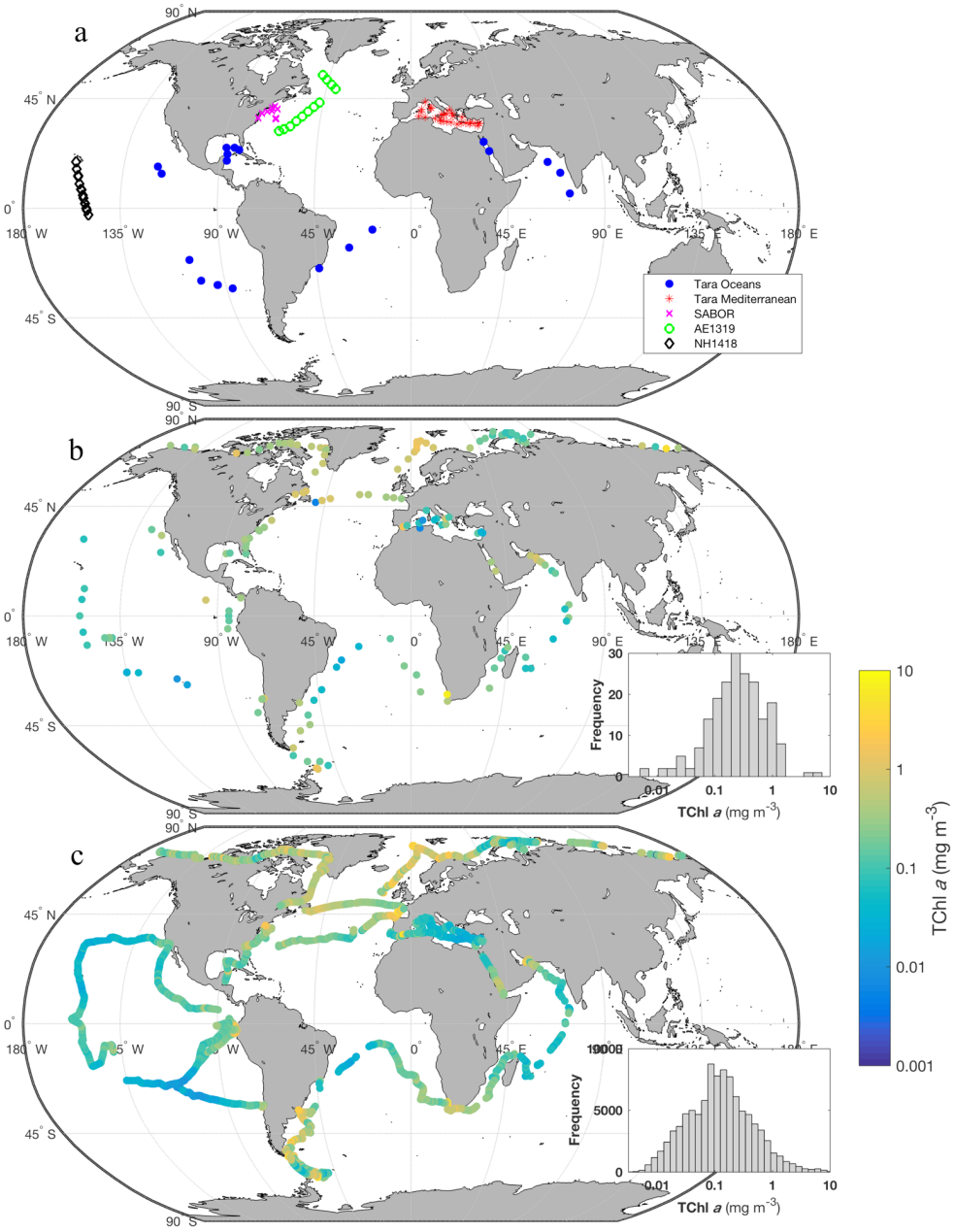

Figure 3.1. Locations of all data used in the present chapter. (a) Locations of coincident $R_{rs}(\lambda)$ and HPLC pigment data measured in situ during five different expeditions used in the inversion analysis (n = 97; see section 3.2.3). (b) Locations of the HPLC pigment data from Tara Expeditions used to calculate covariation relationships between TChl *a* and each of TChl *b,* Chl c_1+c_2, and PPC (n = 196). All data are from the surface. Inset shows the distribution of all TChl *a* data. (c) Locations of $a_p(\lambda)$ spectra measured in situ with an ac-s meter deployed on a flow-through system during the Tara Oceans and

Mediterranean Expeditions (2009 – 2014); n = 96,929. Inset shows the distribution of TChl *a* concentration estimated following the line height method described in *Boss et al.* [2013].

3.2.2. $R_{rs}(\lambda)$ Data and Processing

$R_{rs}(\lambda)$ spectra were calculated from upwelling radiance ($L_u(\lambda)$) and downwelling irradiance ($E_d(\lambda)$) spectra measured with a Profiler II radiometer suite using HyperOCR sensors (Satlantic, Halifax, NS, Canada) deployed with detachable float collar in HyperTSRB ‘‘buoy mode’’ (Figure 3.2). In this configuration, the in-situ radiometers are deployed while the instrument is tethered to the vessel away from the ship shadow and floating at the ocean surface. The sensor that collects downwelling light is above the ocean surface in the air, and the sensor that collects the upwelling light is approximately 0.2 m below the ocean surface. We only keep the $E_d(\lambda)$ and $L_u(\lambda)$ data that fall between the 25th and 75th percentiles to eliminate any outliers before computing $R_{rs}(\lambda)$, which is the desired parameter for ultimately modeling the absorption of phytoplankton pigments via inversion. $R_{rs}(\lambda)$ is defined as

$$R_{rs}(\lambda) = \frac{L_w(\lambda, 0^+)}{E_d(\lambda)}, \tag{3.2}$$

where $L_w(\lambda,0^+)$ is the water-leaving radiance just above the sea surface. $L_w(\lambda,0^+)$ is derived from $L_u(\lambda)$ by first extrapolating $L_u(\lambda)$ from the sensor depth (z = 0.2 m) to just below the sea surface ($L_u(\lambda,z_{0-})$) using

$$L_u(\lambda, z_{0-}) = L_u(\lambda) e^{-K_{Lu}(\lambda)*z}, \tag{3.3}$$

where $K_{Lu}(\lambda)$ is the upwelling attenuation coefficient approximated as

$$K_{Lu}(\lambda) \cong \frac{a_w(\lambda) + a_p(\lambda)}{\mu_u}, \tag{3.4}$$

and μ_u is the average cosine for the upwelling light and is approximated as 0.5.

Absorption by seawater ($a_w(\lambda)$) in the UV and visible is known (Mason et al. 2016; Pope & Fry 1997) and corrected for temperature and salinity (Sullivan et al. 2006), which are measured coincidently. The $a_p(\lambda)$ spectra are measured with the ac-s meter deployed on a flow-through system that records both total and dissolved absorption, and then $a_p(\lambda)$ spectra are calculated by difference (Slade et al. 2010). This method provides $a_p(\lambda)$ measurements that are independent of calibration and instrument drift problems, allowing extended continuous at-sea deployment. However, the method also results in un-calibrated total and dissolved absorption measurements; as a result, equation (3.4) neglects the contribution of CDOM. Additionally, equation (3.4) does not account for inelastic scattering, which will decrease the effective attenuation. There were 31 cases where no coincident $a_p(\lambda)$ data were available for the correction in equation (3.4). For these data we developed an iterative method to estimate the appropriate $a_p(\lambda)$ spectra, as follows: first, the Chl *a* value for a given $R_{rs}(\lambda)$ signal was estimated using the generalized inherent optical properties (GIOP) model (Werdell et al. 2013). Second, the $a_p(\lambda)$ spectrum for the estimated Chl *a* value was defined using the A_{chl} and B_{chl} coefficients determined in this study (see section 3.2.5.1 for details), and used to calculate $K_{Lu}(\lambda)$ (equation (3.4)). Finally, the $K_{Lu}(\lambda)$ value was used in the calculation of $R_{rs}(\lambda)$ (equations (3.2–3.4)). The process was iterated upon until the Chl *a* value determined using GIOP before the attenuation correction was within 1% of the Chl *a* value following the attenuation correction. A self-shading correction was also calculated and applied to the $L_u(\lambda)$ spectra by following the methods in Leathers et al. (2001).

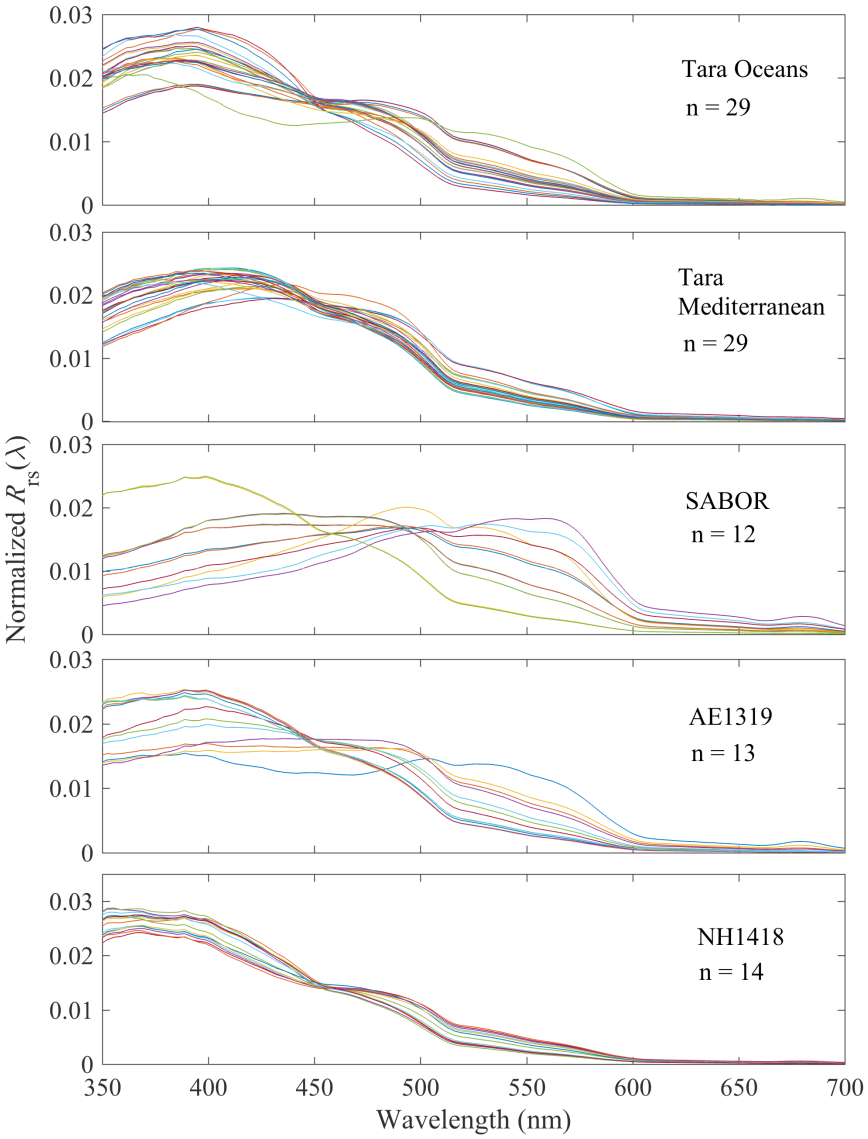

Figure 3.2. $R_{rs}(\lambda)$ spectra measured in situ with a Satlantic HyperTSRB. Each spectrum is normalized to the area under the curve to emphasize differences in spectral shape. See Fig. 3.1a for the locations of each expedition.

The water-leaving radiance exiting the sea surface, $L_w(\lambda,0^+)$, is then calculated as

$$L_w(\lambda, 0^+) = \frac{t\, L_u(\lambda, z_{0-})}{n^2}, \tag{3.5}$$

where n is the refractive index of seawater and t is the radiance transmittance of the surface (assumed to be 1.34 and 0.98, respectively; Mobley 1994).

3.2.2.1. Raman Scattering Correction

Raman scattering by water molecules contributes to the water-leaving radiance ($L_w(\lambda,0^+)$), and can therefore influence $R_{rs}(\lambda)$, particularly in the blue wavelengths in clear waters (Mckinna et al. 2016 and references therein). To account for Raman scattering, the Raman-specific $L_w(\lambda,0^+)$ spectra are computed and subtracted from the original $L_w(\lambda,0^+)$ spectra. Briefly, the steps to complete this are as follows: (1) use a radiation model to estimate the ultraviolet (UV) and visible $E_d(\lambda)$ for the day, time and location on the globe where the in situ $R_{rs}(\lambda)$ data were collected (Gregg & Carder 1990; as coded by Richard Davis, 1998, personal communication); (2) force the measured and modeled $E_d(\lambda)$ spectra to match at the visible wavelengths, and use the corresponding modeled UV values (as the radiometer does not collect data at the necessary UV wavelengths); (3) use the GIOP model (Werdell et al. 2013) to generate inherent optical properties ($a(\lambda)$ and $b_b(\lambda)$; extrapolation used for UV values) for use in calculating $K_{Lu}(\lambda)$ and the downwelling attenuation coefficient ($K_d(\lambda)$); (4) finally, using the calculated K_{Lu} (λ) and $K_d(\lambda)$ values and following methods in Westberry et al. (2013), calculate the Raman-specific $L_w(\lambda,0^+)$ spectra and subtract them from the corresponding original $L_w(\lambda,0^+)$ spectra. The resulting Raman-corrected $L_w(\lambda,0^+)$ spectra are used to calculate $R_{rs}(\lambda)$ (equation (3.2)). The spectral resolution of the final $R_{rs}(\lambda)$ data is approximately 3.35 nm (6 0.05 nm).

3.2.2.2. Correction for Angular Effects

Following the correction for Raman scattering the $R_{rs}(\lambda)$ spectra are normalized to eliminate the angular effect of the sun position in the sky relative to nadir. Following the methods described in Lee et al. (2011), we first use the quasianalytical algorithm (Lee et al. 2002) to estimate $a(\lambda)$ and $b_b(\lambda)$ for a given $R_{rs}(\lambda)$ spectrum. The normalized $R_{rs}(\lambda)$ spectra are then calculated using equations (3.14–3.20) in Lee et

al. (2011), and resulting $R_{rs}(\lambda)$ spectra are used in the inversion algorithm described in section 3.2.3. The absolute percent differences between the original and normalized $R_{rs}(\lambda)$ spectra are 0–5%, 0–6%, and 0.5–9.5% at 440, 490, and 550 nm, respectively. The effect of normalization is small and does not significantly change the outcome during estimation of pigment concentrations; however, we include it as a known effect on the $R_{rs}(\lambda)$ spectra that when accounted for results in spectra that more closely represent the data calculated from satellite information.

3.2.2.3. Inversion of $R_{rs}(\lambda)$ Spectra

To directly model the backscattering and absorption components, we used a model developed by Gordon et al. (1988), where the term $u(\lambda)$ is defined as

$$u(\lambda) \equiv \frac{b_b(\lambda)}{a(\lambda) + b_b(\lambda)}, \tag{3.6}$$

and then used in the quadratic equation

$$r_{rs}(\lambda) = g_1 u(\lambda) + g_2 u(\lambda)^2, \tag{3.7}$$

where $g_1 = 0.0949$ and $g_2 = 0.0794$ are constants computed by Gordon and Brown (1988) and $r_{rs}(\lambda)$ is the remote-sensing reflectance just below the sea surface and can be calculated from $R_{rs}(\lambda)$ using the method from Lee et al. (2002):

$$r_{rs}(\lambda) = \frac{R_{rs}(\lambda)}{0.52 + 1.7R_{rs}(\lambda)}. \tag{3.8}$$

The quadratic formula is used to solve equation (3.7) for $u(\lambda)$ and the one positive solution, denoted $u_{meas}(\lambda)$, is used in a weighted nonlinear least squares inversion to determine the combination of functions representing absorption and backscattering by different constituents in the water that most

closely matches the $u_{meas}(\lambda)$ spectrum. The inversion algorithm iterates to find the best fit while allowing for variation of 31 different parameters (Table 3.3), by minimizing the function

$$\chi^2 = \sum_{i=1}^{60} \left(\frac{u_{\text{meas}}(\lambda_i) - u_{\text{mod}}(\lambda_i)}{u_{\text{std}}(\lambda_i)} \right)^2, \tag{3.9}$$

where the data are summed over the angles 400–600 nm at approximately 3.35 nm resolution (60 total angles), and $u_{mod}(\lambda)$ is the reconstructed spectrum using the spectral components described below. The total backscattering and absorption are broken down into their component spectra, and the expanded version of $u_{mod}(\lambda)$ is defined as

$$u_{\text{mod}}(\lambda) = \frac{b_{bp}(\lambda) + b_{bw}(\lambda)}{a_{\varphi}(\lambda) + a_{CDOM}(\lambda) + a_{NAP}(\lambda) + a_w(\lambda) + b_{bp}(\lambda) + b_{bw}(\lambda)}, \tag{3.10}$$

where $b_{bw}(\lambda)$ is the known backscattering by seawater (Zhang et al. 2009), $a_{\phi}(\lambda)$ is the absorption by phytoplankton, and $a_w(\lambda)$ is the known absorption by seawater (see references in section 3.2.2). The $b_{bp}(\lambda)$ spectrum is defined in terms of the particulate backscattering ratio (b_{bp}/b_p; assumed to be spectrally constant), $a_p(\lambda)$, and particulate attenuation ($c_p(\lambda)$) after Roesler and Boss (2003):

$$b_{\text{bp}}(\lambda) = \frac{b_{\text{bp}}}{b_{\text{p}}} \left(c_{\text{p}}(\lambda) - a_{\text{p}}(\lambda) \right), \tag{3.11}$$

where $b_p(\lambda) = c_p(\lambda) - a_p(\lambda)$; $a_p(\lambda) = a_{\phi}(\lambda) + a_{NAP}(\lambda)$; and $c_p(\lambda)$ is defined as a decreasing power law function:

$$c_{\text{p}}(\lambda) = C_{cp} \left(\frac{\lambda}{\lambda_0} \right)^{-\gamma}. \tag{3.12}$$

The spectra for $a_{NAP}(\lambda)$ and $a_{CDOM}(\lambda)$ are defined as decreasing exponential functions:

$$a_{\mathrm{NAP}}(\lambda) = C_{\mathrm{NAP}} e^{-S_{NAP}(\lambda-\lambda_0)}\,, \tag{3.13}$$

$$a_{\mathrm{CDOM}}(\lambda) = C_{\mathrm{CDOM}} e^{-S_{CDOM}(\lambda-\lambda_0)}\,, \tag{3.14}$$

where λ_0 is set to 400 nm. In equations (3.11–3.14) the following parameters are allowed to vary: b_{bp}/b_p; all of the components in $a_p(\lambda)$ (see equation (3.15) for details on how we define the $a_\phi(\lambda)$ components); the magnitude and slope (C_{cp} and γ, respectively) of $c_p(\lambda)$; and the magnitudes (C_{NAP} and C_{CDOM}) and exponential slopes (S_{NAP} and S_{CDOM}) of $a_{NAP}(\lambda)$ and $a_{CDOM}(\lambda)$.

The spectrum for $a_\phi(\lambda)$ is defined as a linear combination of eight Gaussian functions, rather than a single eigenvector as is usually the case during inversion of $R_{rs}(\lambda)$ spectra into absorbing and scattering components. The individual Gaussian functions ($a_{gaus}(\lambda)$) represent absorption by different phytoplankton pigments or pigment groups, and are defined as:

$$a_{\varphi}(\lambda) = \sum_{i=1}^{8} a_{gaus}(peak_i, \lambda) e^{\left(-0.5\left(\frac{\lambda-peak_i}{\sigma_i}\right)\right)^2}\,, \tag{3.15}$$

with $peak_i$ and σ_i representing the center wavelengths and widths of each Gaussian, respectively, where r is related to the full width half maximum (FWHM) by FWHM = 2.355σ. Although 12 $a_{gaus}(\lambda)$ functions were initially defined, the signal in the red part of the $R_{rs}(\lambda)$ spectra (>600 nm) is relatively low due to the strong absorption by water. We found improved inversion results in terms of reduced median errors when the $R_{rs}(\lambda)$ data were restricted to wavelengths of 400–600 nm; this prevents the inversion from fitting Gaussian functions in the red wavelengths (between 600 and 730 nm) at the cost of poor fitting at the blue end of the spectrum where the $R_{rs}(\lambda)$ signal has lower relative uncertainty. As a result, we use eight Gaussian functions (Table 3.3). A similar finding by Isada et al. (2015) showed that the $R_{rs}(\lambda)$ information >547 nm was dominated by water absorption and therefore not useful for phytoplankton group detection in a Northern Japan bay. We tested several methods for choosing the optimal $peak_i$ and

σ_i values, as evaluated by final correlation and error statistics. These included a derivative analysis and an iterative method in which the $peak_i$ and σ_i values were allowed to change by ±5 nm. However, we ultimately defined the $peak_i$ and σ_i values based on known pigment absorption shapes (see Figure 1, Bricaud et al. 2004) and limited the allowable shift of the $peak_i$ position during inversion to ±1 nm (Table 3.3). The first guess and upper and lower bound values for each parameter in the inversion are based on known values from previous studies of various water types, or from our testing of the inversion algorithm using a range of values (Table 3.3).

The inversion results were improved by normalizing $u_{meas}(\lambda)$–$u_{mod}(\lambda)$ by the uncertainty of $u_{meas}(\lambda)$ at each wavelength (equation (3.9)). We calculated the uncertainty by first calculating $u_{meas}(\lambda)$ for every $E_d(\lambda)$ and $L_u(\lambda)$ spectral pair from a given deployment; for example, a typical deployment from Tara Oceans resulted in roughly 200 spectra from a 3 min time series. We then used the standard deviation of all $u_{meas}(\lambda)$ spectra from one deployment as the uncertainty value ($u_{std}(\lambda)$). These calculations did not include the upper and lower 25th percentiles of $E_d(\lambda)$ and $L_u(\lambda)$ data; their removal is described previously in section 3.2.2.

3.2.3. Estimation of Pigment Concentrations

The concentrations of TChl *a*, TChl *b*, Chl $c_1 + c_2$, and PPC were initially estimated using the relationships between $a_{gaus}(\lambda)$ magnitudes derived from $R_{rs}(\lambda)$ spectra and HPLC pigments (section 3.2.4.1). As a comparison, the concentrations of TChl *b*, Chl $c_1 + c_2$, and PPC were estimated from $R_{rs}(\lambda)$-derived TChl *a* (denoted TChl a_{OC4}) using the covariation relationships among pigments as calculated from a global HPLC data set (section 3.4.2). The TChl a_{OC4} values were calculated using NASA's OC4 algorithm (O'Reilly et al. 2000; see section 3.6 for details). A schematic of the two approaches used to estimate pigment concentrations is provided (Figure 3.3).

Table 3.3. Parameters allowed to vary during the inversion of $R_{rs}(\lambda)$ spectra[a]

Parameter	Units	First guess	Lower bound	Upper bound	Mean ± SD after inversion
C_{NAP}	m^{-1}	0.005	0	0.05	0.004 ± 0.005
S_{NAP}	nm^{-1}	0.011[b]	0.005	0.016	0.013 ± 0.003
C_{CDOM}	m^{-1}	0.1	0.01	0.8	0.047 ± 0.044
S_{CDOM}	nm^{-1}	0.0185	0.005	0.02	0.018 ± 0.002
b_{bp}/b_p		0.01[c]	0.005	0.015	0.007 ± 0.002
C_{cp}	m^{-1}	0.1	0.01	1	0.103 ± 0.070
γ		1	0	1.3	1.156 ± 0.221
$a_{gaus}(\lambda)$ center λ	nm	384[d]	383	385	383.81 ± 0.46
	nm	413[d]	412	414	413.44 ± 0.73
	nm	435[d]	434	436	435.50 ± 0.55
	nm	461[d]	460	462	460.15 ± 0.42
	nm	464[d]	463	465	464.13 ± 0.80
	nm	490[d]	489	491	489.23 ± 0.49
	nm	532[d]	531	533	531.77 ± 0.63
	nm	583[d]	582	584	582.55 ± 0.68
$a_{gaus}(384)$	m^{-1}	0.01	0	0.5	0.014 ± 0.036
$a_{gaus}(413)$	m^{-1}	0.01	0	0.5	0.005 ± 0.006

Table 3.3. continued

$a_{gaus}(435)$	m^{-1}	0.01	0	0.5	0.014 ± 0.015
$a_{gaus}(461)$	m^{-1}	0.01	0	0.5	0.004 ± 0.004
$a_{gaus}(464)$	m^{-1}	0.01	0	0.5	0.007 ± 0.006
$a_{gaus}(490)$	m^{-1}	0.01	0	0.5	0.010 ± 0.007
$a_{gaus}(532)$	m^{-1}	0.01	0	0.5	0.014 ± 0.006
$a_{gaus}(583)$	m^{-1}	0.01	0	0.5	0.022 ± 0.012
$\sigma(384)$	nm	23[d]	22	24	22.81 ± 0.51
$\sigma(413)$	nm	9[d]	8	10	9.86 ± 0.37
$\sigma(435)$	nm	14[d]	13	15	14.77 ± 0.45
$\sigma(461)$	nm	11[d]	10	12	10.22 ± 0.44
$\sigma(464)$	nm	19[d]	18	20	19.85 ± 0.40
$\sigma(490)$	nm	19[d]	18	20	18.24 ± 0.50
$\sigma(532)$	nm	20[d]	19	21	19.63 ± 0.70
$\sigma(583)$	nm	20[d]	19	21	20.80 ± 0.47

[a]See section 3.3 for parameter definitions. n = 97 inversion runs used to calculate mean and standard deviation (SD). b_{bp}/b_p and γ are dimensionless parameters. MATLAB code used for the inversion is available at http://misclab.umeoce.maine.edu/software.php.

[b]*Roesler et al.* (1989)

[c]*Twardowski et al.* (2001)

[d]*Hoepffner and Sathyendranath* (1991); *Bricaud et al.* (2004)

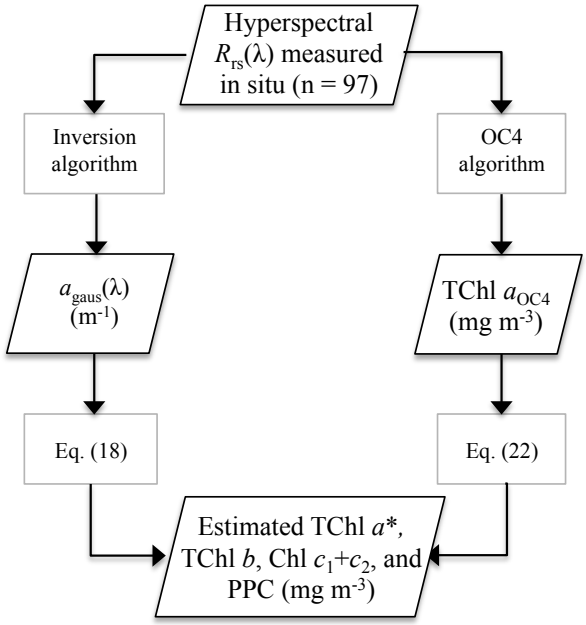

Figure 3.3. Schematic of the data and processing steps used to estimate accessory pigments from hyperspectral $R_{rs}(\lambda)$ data. *Note that TChl *a* in the bottom box is estimated only from $a_{gaus}(\lambda)$; TChl *a* estimated from $R_{rs}(\lambda)$ is computed using the OC4 algorithm (TChl a_{OC4} in middle right box).

3.2.3.1. Pigments Estimated from $a_{gaus}(\lambda)$–HPLC Relationships

Data sets with coincident $a_{gaus}(\lambda)$ derived from inverted $R_{rs}(\lambda)$ and HPLC pigments (n=97; Table 3.1) were compared and used to estimate phytoplankton pigment concentrations: 435 nm and TChl *a*; 464 nm and TChl *b*; 461 nm and Chl $c_1 + c_2$; and 490 nm and PPC (Figure 3.4). We used a nonlinear fit of each $a_{gaus}(\lambda)$ and corresponding HPLC pigment concentration ([pigment_{HPLC}]) to determine the two coefficients (A_{inv} and B_{inv}; Table 3.4) in the equation

$$a_{gaus}(\lambda) = A_{inv}[\text{pigment}_{\text{HPLC}}]^{B_{inv}}, \quad (3.16)$$

by minimizing the cost function

$$\chi^2 = \sum_{i=1}^{97} \left(a_{gaus}(\lambda)_i - A_{inv} * [\text{pigment}_{\text{HPLC}}]_i^{B_{inv}} \right)^2 , \tag{3.17}$$

where there are 97 coincident $a_{gaus}(\lambda)$ and HPLC data points and the function is minimized over the entire data set. The A_{inv} and B_{inv} coefficients obtained from equations (3.16) and (3.17) are then rearranged to solve for the estimated pigment concentration ([pigment$_{inv}$]; mg m^{-3}):

$$[\text{pigment}_{\text{inv}}] = \left(\frac{a_{gaus}}{A_{inv}} \right)^{\frac{1}{B_{inv}}} . \tag{3.18}$$

The estimated pigment concentrations are used to calculate the median percent error (ME):

$$ME = \left(\text{median} \left(\frac{\text{abs}([\text{pigment}_{\text{inv}}] - [\text{pigment}_{\text{HPLC}}])}{[\text{pigment}_{\text{HPLC}}]} \right) \right) * 100 . \tag{3.19}$$

Note that our use of a type 1 nonlinear fit for the data assumes constant uncertainties for all values of $a_{gaus}(\lambda)$ and that HPLC pigment concentrations are error free. Additionally, the relationships determined here are specific to the units of the data used in this study; the A_{inv} values cannot be transferred to estimate pigment concentrations from absorption measurements provided in units other than m^{-1} without proper unit conversion. We also note that throughout this paper, we implicitly assume that any nonlinear coefficients within our reported equations and relationships are normalized, as it is not mathematically possible to exponentiate or compute the logarithm of a dimensionful value. For example, [pigment$_{HPLC}$] in equation (3.16) must be normalized by 1 mg m^{-3} (the units of HPLC pigment data).

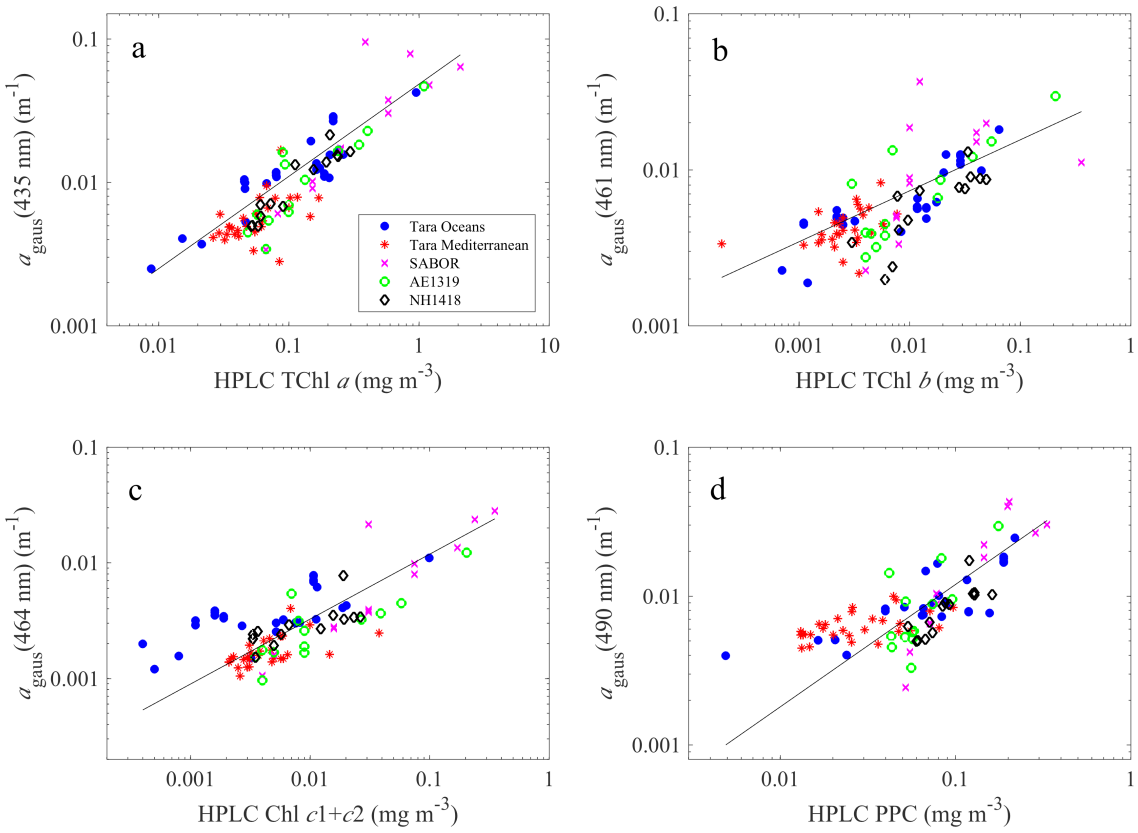

Figure 3.4. Comparisons between HPLC pigment concentrations and $a_{gaus}(\lambda)$ magnitudes. Data are for five different field campaigns as shown in the legend (n = 97). The solid black lines show the type 1 nonlinear fits to the data; see Table 3.4 for coefficients of the best fit lines. Locations of each field campaign are shown in Fig. 3.1a.

3.2.3.2. Pigments Estimated from TChl a_{OC4} Using HPLC Covariation Relationships

As a comparison to the inversion algorithm, pigment concentrations were also estimated using TChl a_{OC4} and covariation relationships between TChl *a* and accessory pigments from HPLC data. The covariation relationships were determined using an HPLC data set spanning various parts of the world ocean (Figures 3.1b and 3.5). To do this, we applied a type 2 linear least squares fit of the log-normalized HPLC data that is weighted by the uncertainties in the data (MATLAB script for the type 2 fit by E. T. Peltzer, MBARI, 2016, http://www.mbari.org/index-of-downloadable-files/). HPLC uncertainties were calculated

using the average percent error for each pigment or pigment group (TChl *a*, TChl *b*, Chl $c_1 + c_2$, and PPC), which is determined by dividing the standard deviation from replicate HPLC samples by the associated measurement. The A_{cov} and B_{cov} coefficients for the covariation pigment relationships (Table 3.4) were determined from the fit between TChl *a* and each of the three accessory pigments ($[\text{pigment}_{HPLC}]$):

$$\log(\text{TChl } a) = \log A_{cov} + B_{cov} * \log([\text{pigment}_{\text{HPLC}}]) \,, \tag{3.20}$$

which is equivalent to

$$\text{TChl } a = A_{cov}[\text{pigment}_{\text{HPLC}}]^{B_{cov}} \,. \tag{3.21}$$

We used the calculated A_{cov} and B_{cov} coefficients from the global HPLC pigment covariation to estimate TChl *b*, Chl c_1+c_2, and PPC from TChl a_{OC4}:

$$[\text{pigment}_{\text{cov}}] = \left(\frac{\text{TChl } a_{\text{OC4}}}{A_{cov}}\right)^{\frac{1}{B_{cov}}} \,. \tag{3.22}$$

The $[\text{pigment}_{cov}]$ concentrations were then used to calculate median error (ME) values using equation (3.19), where $[\text{pigment}_{inv}]$ is replaced by $[\text{pigment}_{cov}]$.

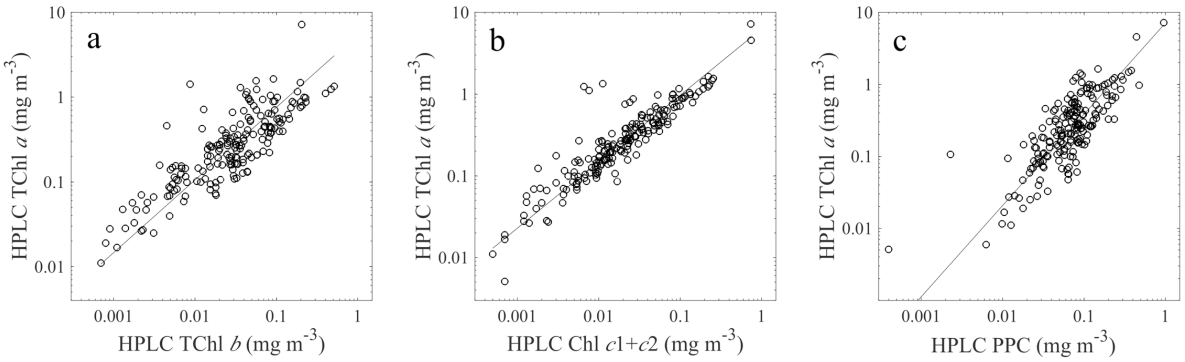

Figure 3.5. Relationships between accessory pigments and TChl *a*. (a) TChl *a* and TChl *b,* (b) TChl *a* and Chl c_1+c_2, and (c) TChl *a* and PPC from HPLC pigment concentrations for the global HPLC dataset (n =

196). The locations of all data points are shown in Fig. 3.1b. The coefficients of the best-fit lines are shown in Table 3.4.

Table 3.4. Coefficients of $a_{gaus}(\lambda)$/HPLC relationships, and HPLC covariation relationships[a]

Wavelength (nm)	HPLC Pigment(s)	from $R_{rs}(\lambda)$ inversion to $a_{gaus}(\lambda)$		from global pigment co-variation	
		A_{inv} (m^{-1})	B_{inv}	A_{cov} (mg m^{-3})	B_{cov}
435	TChl *a*	0.048 ± 0.008	0.643 ± 0.068		
461	Chl c_1+c_2	0.043 ± 0.009	0.561 ± 0.059	6.27 ± 1.08	0.81 ± 0.02
464	TChl *b*	0.033 ± 0.013	0.327 ± 0.074	5.44 ± 1.14	0.86 ± 0.04
490	PPC[b]	0.079 ± 0.024	0.823 ± 0.105	11.10 ± 1.16	1.44 ± 0.06

[a] A_{inv} and B_{inv} calculated using Eq. (3.17); A_{cov} and B_{cov} calculated using Eq. (3.20). Uncertainties were derived using bootstrapping and iterative fitting methods; see section 3.2.4.3. The relationships of pigment covariation compare TChl *a* to HPLC accessory pigments; thus there are no A_{cov} and B_{cov} values for TChl *a*. Note that the wavelengths are only relevant to the inversion analysis, and that B_{inv} and B_{cov} are exponents and therefore unitless.

3.2.3.3. Calculation of Uncertainties in *A* and *B* Coefficients and in Estimated Pigments

To estimate uncertainty values in the A_{inv} and B_{inv} coefficients, we used a bootstrapping method to iteratively subsample all 97 points (with replacement; n = 10,000 iterations) during the fitting between $a_{gaus}(\lambda)$ magnitudes and HPLC pigment concentrations ([$pigment_{HPLC}$]; equation (3.17)). For the A_{cov} and B_{cov} coefficients, the type 2 linear least squares fit described above in section 3.2.4.2 also uses an iterative method to find the best fit between TChl *a* from HPLC and the three accessory pigments ([$pigment_{HPLC}$]). The number of iterations in this case is determined by the fitting routine reaching a user-set convergence limit. Following these iterative methods for both sets of *A* and *B* coefficients, we

calculated the standard deviation of all iterations for each coefficient (A_{inv}, B_{inv}, A_{cov}, and B_{cov}); results are provided as uncertainty values in Table 3.4.

We propagated the *A* and *B* uncertainties using a Monte Carlo approach during the calculation of [$pigment_{inv}$] and [$pigment_{cov}$] to produce uncertainties in estimated pigment values. We multiplied the uncertainty values for each *A* and *B* coefficient (Table 3.4) by a random number (10,000 simulations) pulled from a standard normal distribution. These values are then used to calculate [$pigment_{inv}$] and [$pigment_{cov}$], and the middle 68th percentiles of all values are used to provide uncertainty estimates on the median [$pigment_{inv}$] and [$pigment_{cov}$] estimates (Figures 3.8 and 3.9).

3.2.4. Analysis of Spectral Residuals

3.2.4.1. Calculation of $a_{p\text{-}global}(\lambda)$

To investigate spectral residuals remaining after removal of a Chl *a*-based $a_p(\lambda)$ spectrum, we calculate an ‘‘average global’’ particulate absorption spectrum ($a_{p\text{-}global}(\lambda)$) as a function of Chl *a* concentration. We used a large database of $a_p(\lambda)$ spectra (n = 96,929; Table 3.1; Figure 3.1c) and their corresponding Chl *a* values to determine the relationship between Chl *a* and $a_p(\lambda)$ at each wavelength. The Chl *a* value for each $a_p(\lambda)$ spectrum was calculated using the ac-s line height method and coefficients described in Boss et al. (2013). This data set covers a range of oceanographic conditions and is well representative of the global distribution of chlorophyll values observed by satellite (e.g., Figure 3 in Morel et al. [2007]), with chlorophyll values centered on approximately 0.1 mg m^{-3} (Figure 3.1c inset).

For each wavelength of $a_p(\lambda)$, we regressed the magnitude of $a_p(\lambda)$ against the Chl *a* concentration from the line height calculation; see Figure 2.6 for an example using $a_p(440\ nm)$. We tested both a linear fit of the log-transformed data as well as a nonlinear fit that includes normalization of each $a_p(\lambda)$ value by the associated uncertainty. A linear regression of the log of the variables implicitly assumes that relative uncertainties are constant, which gives more weight to smaller values. However, this is not consistent with what we know about spectrophotometry, where at low values there is an absolute uncertainty

resulting from factors such as instrument resolution and calibration error. To account for uncertainties, we used a nonlinear fit that is normalized by the uncertainties in particulate absorption, denoted $a_{pUNC}(\lambda)$. The $a_{pUNC}(\lambda)$ values were determined by calculating the standard deviation associated with the binning of $a_p(\lambda)$ data to 1 km^2 resolution. We regressed the $a_p(\lambda)$ spectra against Chl *a* by minimizing the following cost function for each wavelength (400–700 nm; 2 nm resolution):

$$\chi^2 = \sum_{\lambda=400}^{700} \left(\frac{a_p(\lambda) - A_{chl}(\lambda) Chl^{Bchl(\lambda)}}{a_{pUNC}(\lambda)} \right)^2 . \tag{3.23}$$

We determined the values of the A_{chl} and B_{chl} coefficients, as well as associated median error values, for each wavelength (Figure 3.7; Table 3.7) similarly to the analysis performed by Bricaud et al. (1995, 1998). We then calculated an $a_{p\text{-}global}(\lambda)$ spectrum that corresponds to each measured $R_{rs}(\lambda)$ by using the A_{chl} and B_{chl} coefficients and TChl a_{OC4} in the equation:

$$a_{\mathrm{p-global}}(\lambda) = A_{chl}(\lambda) TChl\ a^{Bchl(\lambda)} \tag{3.24}$$

Note that as with the relationships derived in section 3.2.4, it is implicitly assumed that the exponentiated value TChl *a* is normalized by 1 mg m^{-3}; with B_{chl} unitless and the units of A_{chl} equal to m^{-1}, the resulting $a_{p\text{-}global}(\lambda)$ will also have units of m^{-1}. We also tested the data for the influence of coastal processes and found no significant change in the fit after removing all data within 10 km of shore (5% of data). Values of median error (ME) for each wavelength were calculated similarly to equation (3.19) by comparing the measured $a_p(\lambda)$ and the $a_{p\text{-}global}(\lambda)$ spectra.

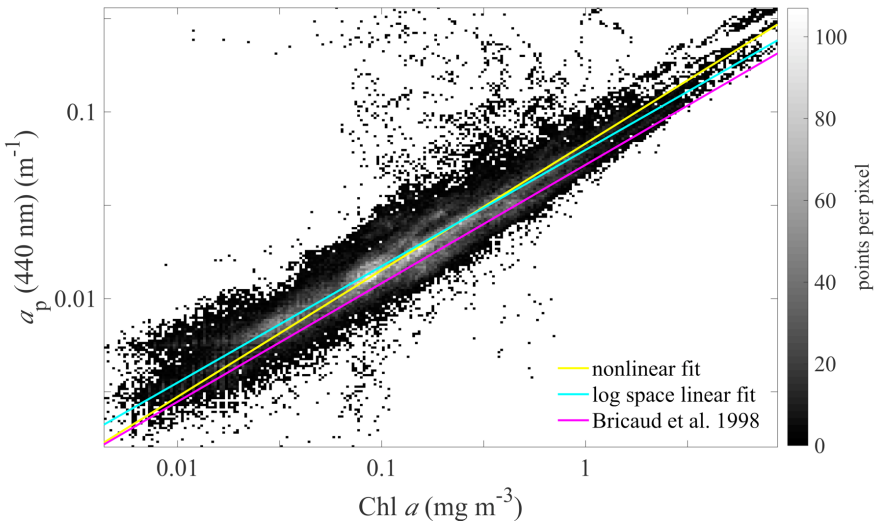

Figure 3.6. Chl *a* estimated using a_p(676 nm) line height regressed against in situ-measured a_p(440 nm). Both the linear fit in log space and the nonlinear fit that also accounts for uncertainties are shown, as well as the fit from *Bricaud et al.* [1998]. The 99.9th percentile of all data are shown for ease of viewing. The gray scale bar represents the number of data points within each 1-km^2 after spatial binning. The outlying points above the best-fit lines are likely caused by a large influence of a_{NAP}(440) to a_p(440).

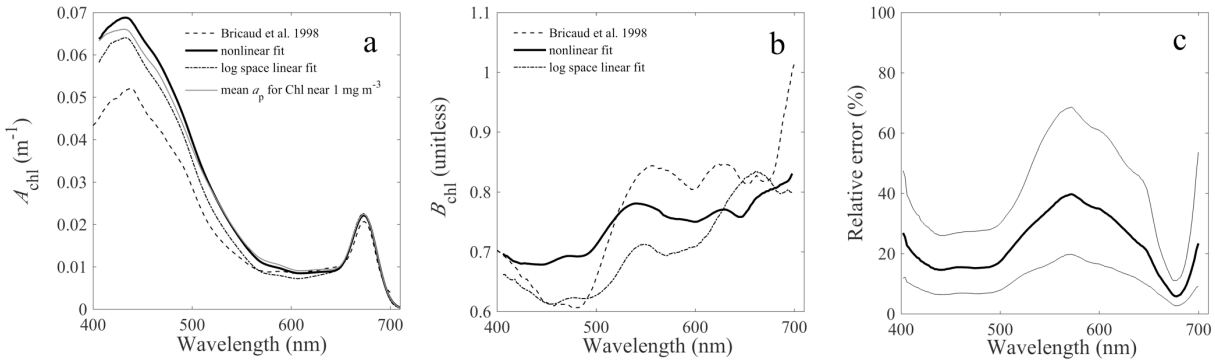

Figure 3.7. Spectral A_{chl} and B_{chl} derived from the relationship: $a_p(\lambda) = A_{chl} * \text{Chl } a^{B_{chl}}$. Values of A_{chl} (panel a) and B_{chl} (panel b) are provided in Table S1. Comparisons between the nonlinear and linear fits are shown, as well as values from *Bricaud et al.* [1998]. Panel (a) shows the $a_p(\lambda)$ spectrum calculated with all values of Chl < 1.05 and > 0.95 mg m^{-3} for comparison with the nonlinear fit, which represents the spectrum for Chl = 1.0 mg m^{-3}. Panel (c) shows spectral values of the relative error associated with the

use of the A_{chl} and B_{chl} coefficients for prediction of $a_p(\lambda)$ from chlorophyll. Thick black line shows the median value (reported as *ME* in Table 3.7) and the bottom and top lines are the 25th and 75th percentiles of the error values, respectively.

3.2.4.2. Calculation and Inversion of Spectral Residuals

Following the calculation of $a_{p\text{-}global}(\lambda)$, we repeated the inversion of $R_{rs}(\lambda)$ with the combined aNAP(λ) and $a_{gaus}(\lambda)$ functions in equation (3.10) replaced by $a_{p\text{-}global}(\lambda)$ to represent the shape of particulate absorption. During the inversion of each spectrum, we set the initial guess for the magnitude of $a_p(\lambda)$ to the magnitude of $a_{p\text{-}global}(\lambda)$ at 440 nm. The resulting $u_{mod}(\lambda)$ (defined in equation (3.10)) is differenced with $u_{meas}(\lambda)$ to calculate a residual spectrum, $u_{resid}(\lambda)$:

$$u_{\mathrm{resid}}(\lambda) = u_{\mathrm{meas}}(\lambda) - u_{\mathrm{mod}}(\lambda)\,. \tag{3.25}$$

The inversion described in section 3.2.3 was repeated for each $u_{resid}(\lambda)$ spectrum (n = 97) and the results were used to compare the $a_{gaus}(\lambda)$ magnitudes with HPLC pigment data, similarly to the analysis described in section 3.2.4. Although the use of $a_{p\text{-}global}(\lambda)$ to calculate $u_{resid}(\lambda)$ removes an average Chl-based absorption spectrum, the residual absorption information that should be contained in $u_{resid}(\lambda)$ can include the influence of Chl *a* as well as any of the accessory pigments, if their relationship to Chl *a* deviates from the global average.

3.3. Results

3.3.1. Pigments Estimated from $a_{gaus}(\lambda)$ and Covariation Relationships

Of the eight $a_{gaus}(\lambda)$ functions used in the inversion, the results of four are presented: $a_{gaus}(435)$, $a_{gaus}(461)$, $a_{gaus}(464)$, and $a_{gaus}(490)$. The $a_{gaus}(384)$ function influences the inversion via the right-hand tail of the Gaussian function that has values ≥400 nm, but because we evaluate the inversion from 400 to 600 nm only, we do not compare $a_{gaus}(384)$ with HPLC pigment data. The remaining three $a_{gaus}(\lambda)$

functions (a_{gaus}(413), a_{gaus}(532), and a_{gaus}(583) did not show significant correlations with any of the pigments or pigment groups we investigated. The a_{gaus}(413), a_{gaus}(532), and a_{gaus}(583) functions represent absorption by TChl *a*, PSC, and Chl $c_1 + c_2$, respectively. Of these, we can estimate TChl *a* and Chl $c_1 + c_2$ using a_{gaus}(435) and a_{gaus}(461), as described above. Using our inversion method, estimates of PSC concentrations cannot be retrieved due to lack of significant correlation between PSC pigments from HPLC and a_{gaus}(532) magnitudes.

We found positive and significant correlations between HPLC pigment concentrations and [pigment$_{inv}$] that range from 0.65 to 0.87 (Spearman's rank correlation coefficient, ρ) and 0.25 to 0.86 (Pearson's linear correlation coefficient, r^2). Median relative errors in prediction of TChl *a*, TChl *b*, Chl $c_1 + c_2$, and PPC using [pigment$_{inv}$] range from 36% to 65% (Table 3.5; Figures 3.8 and 3.9). PPC is estimated with the least error, followed by TChl *a*, Chl $c_1 + c_2$, and TChl *b*. Similar positive and significant correlations were found between pigments estimated from TChl a_{OC4} using covariation relationships ([pigment$_{cov}$]) and HPLC pigment concentrations. The correlation coefficients range from 0.67 to 0.89 (ρ) and 0.51 to 0.96 (r^2). We found increased median error when predicting PPC (52%), the same error for TChl *a* (37%), and lower median errors for Chl $c_1 + c_2$ and TChl *b* (40% and 56%, respectively) for [pigment$_{cov}$] versus [pigment$_{inv}$] (Table 3.5; Figures 3.8 and 3.9).

The increased spread in low values of [pigment$_{inv}$] for all three accessory pigments (Figure 3.8) is also reflected in the spread in the data used to calculate A_{inv} and B_{inv} (Figure 3.4). Chl $c_1 + c_2$ is predicted with the lowest error when using [pigment$_{cov}$], which is also reflected in the low amount of spread around the best-fit line in the data used to calculate A_{cov} and B_{cov} (Figure 3.5b). Several outlying points from the Tara Oceans Expedition (Figure 3.4c) are likely the cause of increased errors in prediction of Chl $c_1 + c_2$ using [pigment$_{inv}$] relative to [pigment$_{cov}$]. Uncertainties in the estimated [pigment$_{inv}$] and [pigment$_{cov}$] values as shown by error bars in Figures 3.8 and 3.9 are relatively small in all cases except for TChl *b* estimated by [pigment$_{inv}$]. This is also reflected in the high ME values for TChl *b* from [pigment$_{inv}$] (Table 3.5).

Table 3.5. Statistics of pigment estimation[a]

Wavelength (nm)	HPLC Pigment(s)	ρ	r^2	*ME* (%)
		from $a_{gaus}(\lambda)$ magnitudes ([pigment$_{inv}$])		
435	TChl *a*	0.87	0.69	37
461	Chl c_1+c_2	0.65	0.86	48
464	TChl *b*	0.75	0.25	65
490	PPC[b]	0.75	0.77	36
		from TChl a_{OC4} ([pigment$_{cov}$])		
	TChl *a*	0.89	0.83	37
	Chl c_1+c_2	0.84	0.96	40
	TChl *b*	0.67	0.51	56
	PPC[b]	0.67	0.70	52

[a]Correlations are Spearman's rank correlation coefficient (ρ) and Pearson's linear correlation coefficient (r^2). *ME* = median error (Eq. 3.19). Wavelengths are only relevant to inversion analysis where $a_{gaus}(\lambda)$ magnitudes are used to estimate pigment concentrations. Note that the statistics comparing TChl *a* with TChl a_{OC4} are a direct comparison; no covariation relationships are used.

[b]PPC = αβ-carotene+zeaxanthin+alloxanthin+diadinoxanthin

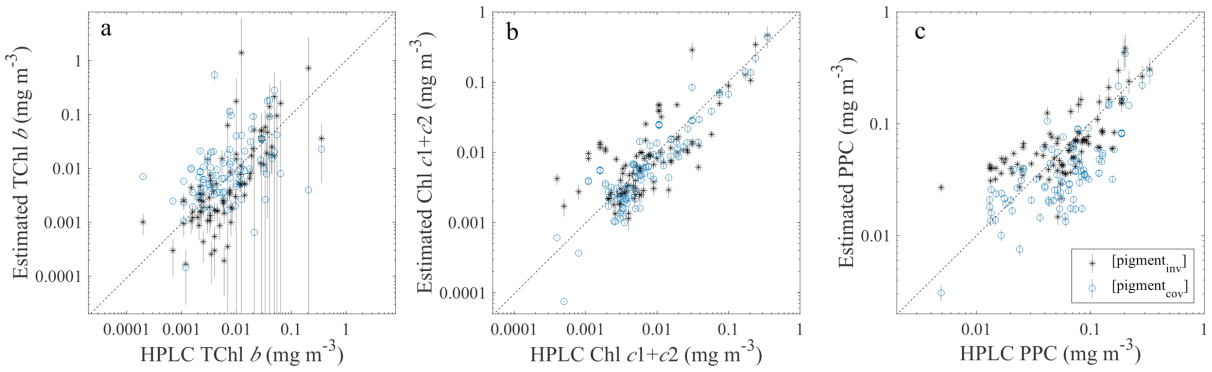

Figure 3.8. Comparisons between HPLC pigment concentrations and pigments. Pigments estimated both from $a_{gaus}(\lambda)$ ([pigment$_{inv}$], black "*") and TChl a_{OC4} ([pigment$_{cov}$], blue open circles) (y-axes). Dotted lines show one-to-one relationships; n = 97. Vertical lines show the first standard deviation of estimated pigments; see section 3.2.4.3 for more details.

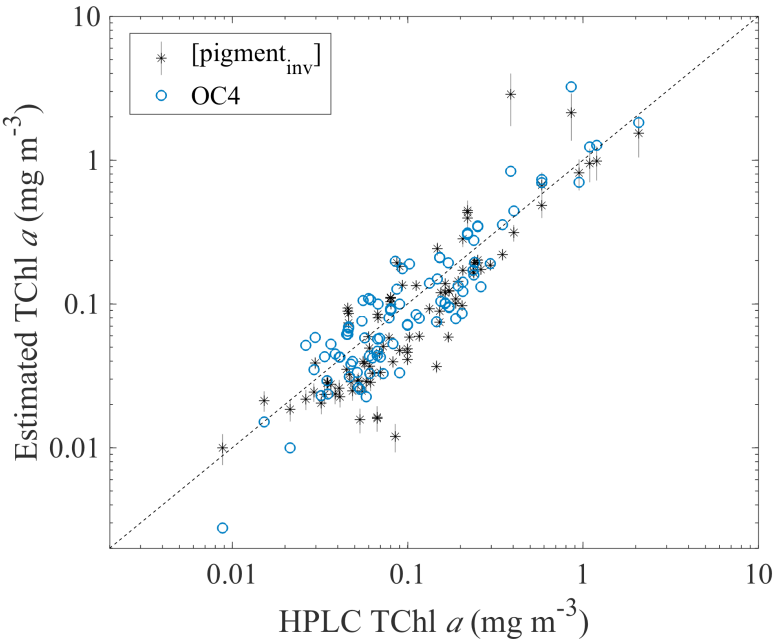

Figure 3.9. Comparison of TChl *a* computed from spectra with HPLC TChl *a*. Dotted line shows the one-to-one relationship; n = 97. Vertical lines on the [pigment$_{inv}$] values show the first standard deviation of estimated pigments; see section 3.2.4.3 for more details. Uncertainties are not provided for the TChl a_{OC4} values as we do not have information on the uncertainty of parameters used in the OC4 algorithm.

3.3.2. Pigment Ratios

As a test of consistency with previous studies, we compared the distributions of pigment ratios determined using HPLC pigments, the [pigment$_{inv}$] concentrations from $a_{gaus}(\lambda)$ magnitudes, and the [pigment$_{cov}$] concentrations from TChl a_{OC4} using global covariation relationships (Figure 3.10). The distribution of TChl *b*:TChl *a* from TChl a_{OC4} ([pigment$_{cov}$]) are shifted slightly higher than TChl *b*:TChl *a* from the other two data sets (Figure 3.10a). The distributions of Chl $c_1 + c_2$:TChl *a* from HPLC and both [pigment$_{inv}$] and [pigment$_{cov}$] follow a similar pattern (Figure 3.10b). The mode values of the PPC:TChl *a* distribution from both [pigment$_{inv}$] and [pigment$_{cov}$] are slightly higher than that from the HPLC pigment data (Figure 3.10c). The distributions of all values are within the ranges of previously published values (see Table 3 in Chase et al. 2013).

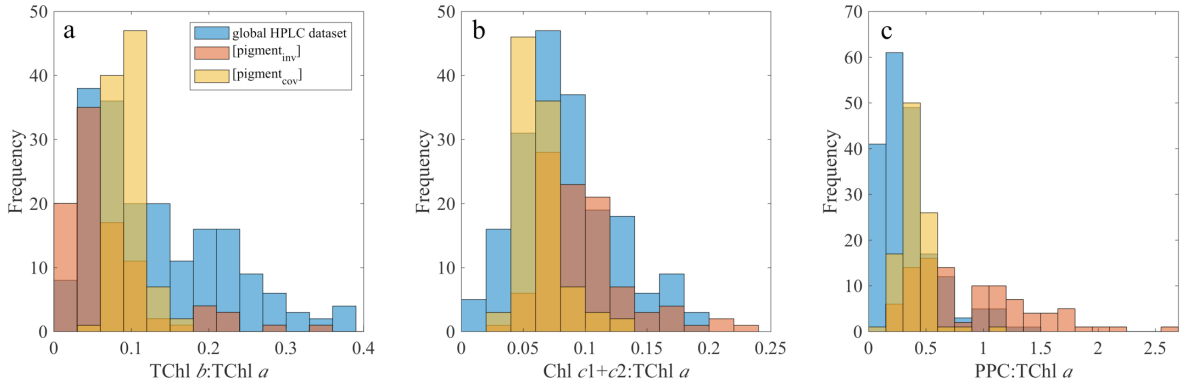

Figure 3.10. Frequency distributions of pigment ratios. Panels show (a) TChl *b*:TChl *a*, (b) Chl c_1+c_2:TChl *a,* and (c) PPC:TChl *a* determined from the global HPLC dataset (n = 196, blue), the pigment concentrations estimated from $a_{gaus}(\lambda)$ ([pigment$_{inv}$]; n = 97, orange), and the pigment concentrations estimated from TChl a_{OC4} ([pigment$_{cov}$]; n = 97, yellow). The details of all datasets are presented in Table 3.1. Three high [pigment$_{inv}$] data points are excluded for ease of viewing, with ratio values of 0.48 and 0.75 in (a) and 4.2 in (c).

3.3.3. Spectral Residuals

The values of spectral A_{chl} and B_{chl} are shown in Figures 3.7a and 3.7b and provided in supporting information Table 3.7. Deviations from the values reported in Bricaud et al. (1998) are likely the result of both an increased number of data points at the lower end of the Chl *a* value range in our data set, and the nonlinear fit used in our calculation of A_{chl} and B_{chl}. Note that the difference in scattering correction of the $a_p(\lambda)$ data (our nonlinear correction versus a spectrally constant scattering correction by Bricaud et al. 1998) cannot explain the difference as ours is expected to remove more scattering out of absorption in the blue wavelengths, which would preferentially decrease the $a_p(\lambda)$ values in that region of the spectrum. The relative error values are highest in the region of the spectrum where absorption is generally the lowest (Figure 3.7c).

The inversion of $u_{resid}(\lambda)$ spectra (defined in section 3.2.5.2) was conducted to allow for the extraction of pigment information contained in any spectral features in $a_p(\lambda)$ that deviate from $a_{p\text{-}global}(\lambda)$. We found no significant correlations between the $a_{gaus}(\lambda)$ magnitudes following the inversion of $u_{resid}(\lambda)$ spectra and HPLC pigment concentrations (not shown). In the inversion analysis, $a_{CDOM}(\lambda)$ and $b_b(\lambda)$ are allowed to vary during the inversion as they were not measured in situ and so are not known *a priori*. Because of this, during the inversion analysis of $u_{resid}(\lambda)$ the inverted $a_{CDOM}(\lambda)$ and $b_b(\lambda)$ functions compensated for the differences between the measured $a_p(\lambda)$ and $a_{p\text{-}global}(\lambda)$ spectra, which prevented any extraction of pigment information beyond the average global relationships between TChl *a* and accessory pigments.

3.4. Discussion

Changes in phytoplankton community composition are known to be correlated with changes in Chl *a* concentration, as smaller cells are linked to regenerated production and dominating in oligotrophic waters (Chl $a < 0.1$ mg m^{-3}), and larger cells (namely diatoms) are associated with the ‘‘new’’ production of higher biomass regions (Chisholm 1992; Eppley & Peterson 1979). Surface Chl *a* values have also been

related to phytoplankton size classes determined by HPLC diagnostic pigments (Ras et al. 2008; Uitz et al. 2006). Understanding the relationships between hyperspectral reflectance measurements and underlying properties is critical for developing algorithms that can be used to study phytoplankton community composition on a global scale. In the work presented here, we explored the limits of using global, hyperspectral $R_{rs}(\lambda)$ data to estimate phytoplankton pigments, which in turn can be used to help define phytoplankton composition. We also used global covariation relationships between TChl *a* and HPLC pigments to estimate accessory pigment concentrations from $R_{rs}(\lambda)$-derived Chl *a*; these estimates provide a bench-mark against which any other method for accessory pigment estimation should be tested.

3.4.1. Estimation of Pigment Concentrations

In the comparison of our results to previous studies, we consider both the median relative errors resulting from the spectral inversion algorithm and from the pigment covariation method (Table 3.5). The median relative errors of estimated pigments from our study are higher than those reported in Wang et al. (2016), but several key differences are noted. First, their study compared only $a_{gaus}(\lambda)$ magnitudes from the decomposition of $a_p(\lambda)$ and $R_{rs}(\lambda)$ spectra (i.e., no HPLC data were available), and therefore the method was not evaluated using measured pigment concentrations. Second, their data set consisted of measurements made in fresh inland waters with Chl *a* values ranging from 10 to 1,377 $\mu g\ L^{-1}$. Additionally, the high reflectance values in the red region (Figure 4 in Wang et al. 2016) allow their use of the entire visible spectrum for the inversion analysis; this is not possible with open ocean $R_{rs}(\lambda)$ spectra that have a much lower signal above 600 nm (e.g., Figure 3.2, this study). When compared with our study, median errors in pigments estimated by Bracher et al. (2015) are lower for estimation of TChl *a*, Chl c_1+c_2, and PPC. The study by Bracher et al. (2015) did not include TChl *b*, as a result of a high percentage of samples with no detected TChl *b* (24% and 34% for satellite and field samples, respectively; see supplement of Bracher et al. 2015). The mean absolute percent error values found by

Pan et al. (2010) are slightly lower for TChl *a*, TChl *b*, and TChl *c* (which includes Chl c_3), and higher for PPC (in their study each PPC pigment is estimated individually). Both the Pan et al. (2010) and Bracher et al. (2015) studies were regionally adapted for the United States northeast coast and the Atlantic Ocean, respectively. Although the median error values we calculated were slightly higher than the values from these two previous studies, our method is theoretically applicable at a global scale.

3.4.2. Pigment Covariation

The study by Pan et al. (2010) showed high correlation coefficients between TChl *a* and TChl *c*, and TChl *a* and each of the PPC pigments used in our study except for zeaxanthin. The global covariation between Chl *a* and phytoplankton accessory pigments was also observed previously by Trees et al. (2000), with a similar slope value (0.93) of the log-linear relationship between Chl *a* and total summed accessory pigments (both in mg m^{-3}) to what we observed in the global HPLC data (column of B_{cov} values in Table 3.4; mean slope value of 1.04). A previous study by Uitz et al. (2015) found strong correlations between the first and second EOF mode (describing a combined 95% of the variance) of $R_{rs}(\lambda)$ spectra and several pigments (Chl *a*, zeaxanthin, 190-hexanoyloxyfucoxanthin, and fucoxanthin). Bracher et al. (2015) showed that the second EOF is the most important for the prediction of several pigments and pigment groups: TChl *a*, monovinyl Chl *a* (MVChl *a*), PSC, and PPC. In the case of both of these previous studies, the prediction of Chl *a* and other pigments from the same EOF modes further supports the covariation of Chl *a* and accessory pigments in ocean waters.

Our previous work to estimate pigments from $a_p(\lambda)$ spectra (Chase et al. 2013) did not examine the use of pigment covariation for estimating accessory pigments; however, we can compare the results of that study with the statistics generated here for estimated pigments. There are similar or slightly lower errors using $a_{gaus}(\lambda)$ from decomposition of ac-s spectra as reported in Chase et al. (2013) compared to the median errors in the present work, with the exception of PPC, which is estimated with lower error using the inversion algorithm in the current study. The lower errors in estimating chlorophylls *a*, *b*, and *c*

directly from $a_p(\lambda)$ is likely made possible by the distinct absorption peak positions of the chlorophylls compared to other pigments in the red part of the visible spectrum, where there is sufficient information in $a_p(\lambda)$ but not $R_{rs}(\lambda)$ spectra. One source of uncertainty in pigment covariation relationships is the spread around the best-fit line that results from natural variations in phytoplankton pigment composition. In addition, pigment covariation relationships are limited by the accuracy of the TChl *a* that is subsequently used to estimate the other accessory pigments. Uncertainties in the estimation of TChl *a* are variable depending on the method used (e.g., HPLC analysis, spectral band ratio, ac-s line height) and should be taken into consideration when using TChl *a* and pigment covariation parameters to estimate accessory pigments.

3.4.3. Spectral Resolution Considerations

Our use of spectral inversion and Gaussian functions to estimate phytoplankton pigments from hyper-spectral $R_{rs}(\lambda)$ is an analytical approach that attempts to exploit any influence of pigment absorption on the spectral shape of $R_{rs}(\lambda)$. One appeal of such an approach is that it does not require training data sets or any previous knowledge of the water of interest, thereby making it more robust in an ocean experiencing large-scale climate-driven changes. In addition, it eliminates the need to define the shape of $a_\phi(\lambda)$. However, there are still assumptions made when defining the shapes of $a_{NAP}(\lambda)$ and $a_{CDOM}(\lambda)$ spectra used in the inversion algorithm. This could present a problem if, for example, $a_{NAP}(\lambda)$ spectra are not exactly exponential, which has been seen in previous studies (Babin & Stramski, 2004; Estapa et al. 2012; Iturriaga & Siegel, 1989). Uncertainties in the prescribed spectral shapes, as well as in the Raman correction (section 3.2.2.1), are potential sources of error in the inversion and subsequent magnitudes of $a_{gaus}(\lambda)$ spectra.

The peak locations of the functions representing TChl *b* and Chl $c_1 + c_2$ absorption in our analysis (464 nm and 461 nm, respectively) are located only three nanometers apart; however, the defined widths of the two Gaussian functions are different (Table 3.3). Both the peak locations and widths are based on

laboratory-measured pigment absorption (Bricaud et al. 2004). The method of spectral decomposition into multiple Gaussian functions uses both the peak location and width, which can allow for the separation of absorption features that are close spectrally. Additionally, hyperspectral data provide the capability for inversion methods using Gaussian decomposition, while multispectral data do. Previous studies have used spectral derivative methods (Lee et al. 2007; Wolanin et al. 2016) to determine the optimal wavelength locations needed to maximize the information extracted from $R_{rs}(\lambda)$ spectra. Wolanin et al. (2016) concluded that hyperspectral data is the most effective for discriminating between three phytoplankton groups (diatoms, coccolithophores, and cyanobacteria). The derivative analysis by Lee et al. (2007) showed that a multispectral approach could be used to identify the location of local extremes or inflections. However, an inversion method using Gaussian decomposition can be used to estimate pigments that may be influencing the reflectance signal without creating local features that can be enhanced by derivative analysis.

3.4.4. Spectral Residuals Analysis

We suspect that factors other than phytoplankton accessory pigments are driving the deviation of $a_p(\lambda)$ spectra from the $a_{p\text{-}global}(\lambda)$ expected for a given Chl *a* value. For example, as mentioned above, $a_{NAP}(\lambda)$ spectra may not be exactly exponential. Reflectance anomalies calculated using multispectral satellite data by Huot and Antoine (2016) showed that several $R_{rs}(\lambda)$ anomalies were well correlated with either $a_{CDOM}(\lambda)$ anomalies or $b_{bp}(\lambda)$ estimates. However, some of these anomalies (namely those determined using the ratio of R_{rs}(488 nm) to other bands) were not well correlated with any other satellite ocean color products, but rather are likely due to some combined variation of $a_{CDOM}(\lambda)$, $a_{\phi}(\lambda)$, and/or $b_{bp}(\lambda)$. Of these three, $a_{\phi}(\lambda)$ contains the most spectral variability, due to the presence of different accessory pigments. This supports the possibility of using hyperspectral reflectance residuals to detect phytoplankton accessory pigment signatures; however, more work is needed to understand how to best

use the residuals of hyperspectral $R_{rs}(\lambda)$ data for obtaining information on phytoplankton community composition.

3.5. Conclusions

Direct estimation of phytoplankton pigments from $R_{rs}(\lambda)$ spectra has the appeal of application to a wide range of ocean waters; this is especially desirable for use with satellite remote-sensing data that covers much of the globe. Here, we have estimated phytoplankton pigments from in situ hyperspectral $R_{rs}(\lambda)$ data that spans various oceanic water types. We determined that several phytoplankton accessory pigments can be predicted via an inversion algorithm with median errors ranging from 36% to 65%. Additionally, we calculated similar predictive capabilities for pigments when a global relationship of pigment covariation is applied to TChl *a* concentrations derived from $R_{rs}(\lambda)$ spectra. The median error values provided in this study are a crucial component in efforts to estimate phytoplankton groups from optical and ultimately satellite data, and a lack of uncertainty information has been an area identified as a gap in the current research (Bracher et al. 2017). However, we note the similar values in the predictive capabilities of the two methods, and that there will be variability in the predicted pigment uncertainties due to factors such as the algorithm used to estimate Chl *a* from $R_{rs}(\lambda)$. The high covariation of accessory pigments with TChl *a*, on a global scale, should be carefully considered when using algorithms that are designed to estimate phytoplankton accessory pigments from $R_{rs}(\lambda)$ spectra; we propose that covariation relationships should be used as a bench-mark against which to evaluate the utility of novel methods. Although changes in phytoplankton biomass may be correlated with phytoplankton community composition, extracting knowledge on community composition beyond information based on TChl *a* and its covarying parameters is not trivial. While the analysis of spectral residuals reported here did not show the utility of residuals from hyperspectral $R_{rs}(\lambda)$ to estimate pigment concentrations, this approach should be explored further as one way to move beyond the information provided from

the covariation of accessory pigments with TChl *a*. Finally, automated phytoplankton imagery and molecular data are rapidly becoming more available and will be important for comparison with both HPLC data and optical spectra. Future research should include analysis and validation work to compare HPLC pigment information with phytoplankton community composition from other data such as molecular information or imagery, thus connecting the estimation of pigments with phytoplankton community in a framework with known uncertainties.

3.6. Comparison of Chlorophyll Algorithms

We tested both NASA's current OCI algorithm (https://oceancolor. gsfc.nasa.gov/atbd/chlor_a/) and the standard OC4 algorithm (O'Reilly et al. 2000) when calculating the $R_{rs}(\lambda)$-derived estimates of TChl *a*. The OCI algorithm is a combination of the standard OC4 band ratio algorithm combined with the Color Index (CI) from Hu et al. (2012). At Chl *a* values above 0.2 mg m^{-3} OC4 is used, at values below 0.15 mg m^{-3} CI is used, and a weighted combination of the two algorithms is used for values in between. We found that the comparison between $R_{rs}(\lambda)$-derived TChl *a* and HPLC TChl *a* was slightly improved when using OC4 (Figure 3.11 and Table 3.6); as a result, we opted to use the OC4 algorithm through the study. Note that we also tested the use of Chl *a* versus TChl *a* and the statistics and results are similar.

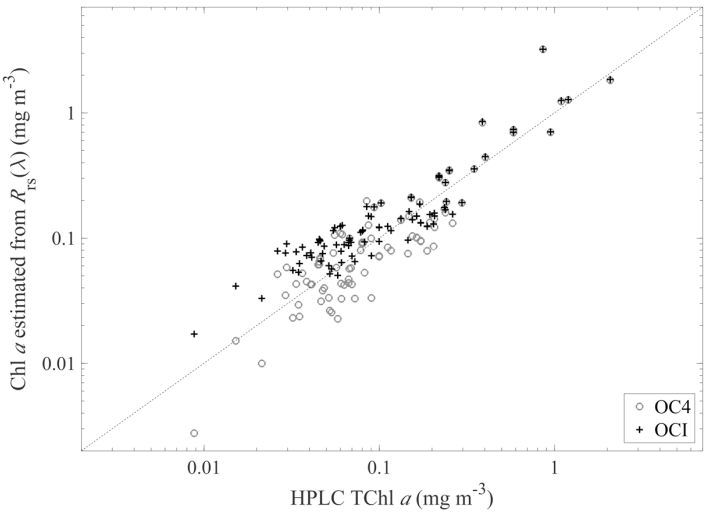

Figure 3.11. Chl *a* estimated from R_{rs}(λ) spectra. Chl *a* estimated both using OC4 (grey open circles) and OCI (black plus symbols) compared to TChl *a* from HPLC analysis. R_{rs}(λ) spectra are from the same five datasets used in the inversion analysis (n = 97; Table 3.1; Fig. 3.2).

Table 3.6. Statistics calculated from the comparison of the OC algorithms with HPLC TChl *a*[a]

Chl *a* algorithm	*ρ*	r^2	*ME* (%)	*MeanE* (%)	*MBE* (%)	*RMSE*
OC4	0.86	0.8	35	39	4	0.25
OCI	0.89	0.8	39	53	43	0.25

3.7. Supplementary Information

This supporting information provides the caption for the table of values for A_{chl} and B_{chl} used in the calculation of $a_{p\text{-}global}$(λ) (Eq. 3.24). Median error (*ME*) values are also provided as a metric of uncertainty in the a_p(λ) values predicted from Chl *a.*

Table 3.7. Spectral coefficients and errors calculated for the Chl *a* and $a_p(\lambda)$ global relationship. Values of A_{chl} and B_{chl} are calculated using a nonlinear fit between Chl *a* and $a_p(\lambda)$ for each wavelength: $a_p(\lambda) = A_{chl} * \text{Chl } a^{B_{chl}}$. *ME* = median relative error.

Wavelength (nm)	A_{chl} (m^{-1})	B_{chl} (unitless)	*ME (%)*
406	0.064	0.697	22.7
408	0.064	0.693	21.9
410	0.065	0.693	20.9
412	0.066	0.69	20.1
414	0.066	0.687	19.4
416	0.067	0.685	18.7
418	0.067	0.683	18.4
420	0.067	0.683	17.5
422	0.068	0.681	17.2
424	0.068	0.68	16.7
426	0.068	0.68	16.3
428	0.069	0.681	16
430	0.069	0.68	15.4
432	0.069	0.68	15.2
434	0.069	0.679	14.8
436	0.069	0.679	14.8
438	0.068	0.679	14.6
440	0.068	0.679	14.6
442	0.067	0.679	14.6
444	0.066	0.678	14.7
446	0.065	0.678	14.9
448	0.065	0.679	14.9
450	0.064	0.68	15.1
452	0.063	0.681	15.3
454	0.063	0.682	15.3
456	0.062	0.684	15.5
458	0.061	0.685	15.5
460	0.061	0.687	15.5
462	0.06	0.688	15.5
464	0.059	0.689	15.4
466	0.058	0.691	15.4
468	0.058	0.692	15.4
470	0.057	0.693	15.3
472	0.056	0.693	15.2
474	0.055	0.693	15.1

Table 3.7. continued

476	0.054	0.693	15.2
478	0.053	0.692	15.2
480	0.052	0.692	15.3
482	0.051	0.692	15.3
484	0.05	0.692	15.3
486	0.049	0.693	15.4
488	0.048	0.693	15.4
490	0.046	0.695	15.6
492	0.045	0.697	15.8
494	0.044	0.699	16.1
496	0.042	0.703	16.4
498	0.041	0.707	16.8
500	0.04	0.711	17.4
502	0.038	0.715	18
504	0.037	0.72	18.8
506	0.036	0.725	19.6
508	0.035	0.73	20.4
510	0.033	0.735	21.1
512	0.032	0.74	22
514	0.031	0.745	22.8
516	0.03	0.749	23.6
518	0.029	0.753	24.3
520	0.028	0.757	25.1
522	0.027	0.762	25.9
524	0.026	0.765	26.6
526	0.025	0.769	27.4
528	0.024	0.772	28.1
530	0.024	0.774	28.7
532	0.023	0.776	29.4
534	0.022	0.778	30
536	0.021	0.779	30.6
538	0.02	0.78	31.2
540	0.02	0.781	31.8
542	0.019	0.78	32.4
544	0.018	0.78	33
546	0.017	0.78	33.7
548	0.017	0.779	34.2
550	0.016	0.778	34.8
552	0.015	0.777	35.4
554	0.015	0.776	36
556	0.014	0.775	36.8
558	0.014	0.773	37.3

Table 3.7. continued

560	0.013	0.772	38
562	0.013	0.77	38.5
564	0.012	0.768	38.8
566	0.012	0.765	39.1
568	0.011	0.761	39.3
570	0.011	0.76	39.6
572	0.011	0.758	39.7
574	0.011	0.757	39.5
576	0.01	0.756	39.1
578	0.01	0.756	38.6
580	0.01	0.755	38.2
582	0.01	0.754	37.7
584	0.01	0.755	37.1
586	0.01	0.754	36.6
588	0.01	0.754	36.2
590	0.01	0.753	35.9
592	0.009	0.753	35.6
594	0.009	0.752	35.4
596	0.009	0.751	35.1
598	0.009	0.75	35
600	0.009	0.75	34.9
602	0.009	0.75	34.5
604	0.009	0.751	34
606	0.008	0.753	33.5
608	0.008	0.755	33.1
610	0.008	0.756	32.6
612	0.008	0.758	32.2
614	0.008	0.76	31.5
616	0.009	0.762	30.9
618	0.009	0.765	30.3
620	0.009	0.766	29.7
622	0.009	0.768	29
624	0.009	0.769	28.4
626	0.009	0.77	27.8
628	0.009	0.77	27.1
630	0.009	0.771	26.4
632	0.009	0.77	25.8
634	0.009	0.768	25
636	0.009	0.767	24.5
638	0.009	0.764	23.7
640	0.009	0.762	23.3
642	0.009	0.76	22.7

Table 3.7. continued

644	0.009	0.758	22.3
646	0.009	0.758	21.7
648	0.01	0.759	20.9
650	0.01	0.762	19.5
652	0.011	0.767	18.3
654	0.011	0.771	16.8
656	0.012	0.777	15.4
658	0.014	0.781	14
660	0.015	0.787	12.8
662	0.016	0.79	11.9
664	0.018	0.792	11
666	0.019	0.795	10
668	0.021	0.798	9.2
670	0.021	0.799	8.2
672	0.022	0.801	7.2
674	0.022	0.801	6.4
676	0.022	0.804	5.9
678	0.021	0.804	5.8
680	0.02	0.806	6.1
682	0.018	0.807	6.4
684	0.016	0.809	7.3
686	0.014	0.812	8.3
688	0.012	0.813	9.6
690	0.01	0.815	11.4
692	0.008	0.816	13.3
694	0.006	0.818	15.6
696	0.005	0.823	18
698	0.003	0.831	21.2
700	0.003	0.835	23.4

CHAPTER 4

EVALUATION OF DIAGNOSTIC PIGMENTS TO ESTIMATE PHYTOPLANKTON SIZE CLASSES

4.1. Introduction

Phytoplankton community composition, i.e., the assemblage of different species present in the ocean, is variable in both time and space. Phytoplankton species span multiple taxonomic kingdoms and phyla, and several orders of magnitude in cell size (< 1 to >1000 µm). Functional traits of phytoplankton communities have been applied as a way to move beyond bulk community parameters (e.g., chlorophyll-*a* concentration [Chl *a*]), while reducing the need for collecting information on individual species to a manageable and meaningful number of groups (Litchman et al. 2010; Le Quere et al. 2005). Phytoplankton cell size is an important trait that influences nutrient uptake, growth rates, sinking rates, motility, and interactions with grazers (Litchman and Klausmeier 2008; Sommer et al. 2017). Three size classes of phytoplankton are commonly used, originally defined by Sieburth et al. (1978): picoplankton (0.2–2 μm), nanoplankton (2–20 μm), and microplankton (20–200 μm); most subsequent literature uses < 2 μm to define picoplankton and > 20 μm to define microplankton. For brevity in this paper, we indicate the three size classes of phytoplankton using the terms pico-, nano-, and microplankton without "phyto" included. These size classes follow from the observation of Sheldon et al. (1972) that the biomass of phytoplankton in power-law increasing size bins is approximately equal. There are some expected relationships between major phytoplankton taxonomic groups and the three size classes. For example, cyanobacteria from the genera *Synechococcus* and *Prochlorococcus* generally fall under the operational definition of picoplankton (Chisholm 1992) (although we note that *Trichodesmium* colonies can reach microplankton size, or larger in some tropical and subtropical regions). Most cryptophytes are found in a relatively narrow range of sizes within the nanoplankton (Roy et al. 2011; Tomas 1997). However, in some phytoplankton groups, such as diatoms and dinoflagellates, cell size spans a broad range from several to hundreds of micrometers. Ultimately, knowledge of the distribution of

phytoplankton cell sizes, often termed "phytoplankton size classes" (PSCs) provides insights into ecosystem processes (e.g., carbon flux, nutrient cycling) and food web dynamics (namely grazer interactions and preferences) that are of importance in the ocean.

Efforts to observe and monitor the size structure of phytoplankton communities in the open ocean are limited by the challenges of in situ data collection. Satellite data make possible the efforts to quantify PSCs on regional or global ocean scales (e.g., Brewin et al. 2010; Hirata et al. 2011; Li et al. 2013; Roy et al. 2013; Kostadinov et al. 2009; also see Mouw et al. 2017 for a review of recent literature). Development and validation of satellite algorithms requires extensive in situ data, and in the case of PSCs, direct measurements of physically separated size fractions of phytoplankton biomass (as much as is possible with size-fractionated filtration) are limited. Vidussi et al. (2001) proposed the use of phytoplankton pigment information as measured by High Performance Liquid Chromatography (HPLC) analysis as an alternative to labor-intensive size fractionation and microscopy approaches. The motivation for this approach, which builds off a method proposed by Claustre (1994) to estimate the relative abundance of microplankton, is that in addition to Chl *a*, phytoplankton contain various photosynthetic and photoprotective accessory pigments that are used for visible light absorption. These different accessory pigments are associated with broad taxonomic phytoplankton groups, which are a result of multiple endosymbiosis events during the evolution of different eukaryotic algal lineages, as well as the evolution of photosynthetically active cyanobacteria (Jeffrey et al. 1997). Pigments contained within phytoplankton taxonomic groups are in turn assumed to be associated with the three size classes of pico-, nano-, and microplankton.

Vidussi et al. (2001) used seven diagnostic accessory pigments (fucoxanthin (Fuco), peridinin (Peri), 19'-hexanoyloxyfucoxanthin (Hexa), 19'-butanoyloxyfucoxanthin (Buta), alloxanthin (Allo), total chlorophyll-*b* (TChlb), and zeaxanthin (Zea)) to derive phytoplankton size classes. Generalizations must be made

when assigning accessory pigments to phytoplankton taxonomic groups, as well as groups to each of the three size classes. The method of using diagnostic accessory pigments, referred to as the Diagnostic Pigment Analysis (DPA) throughout this paper, was further developed by Uitz et al. (2006), who applied weighting coefficients to the same seven diagnostic pigments used by Vidussi et al. (2001) to describe the contribution of each PSC to total Chl *a*. The original assignment of pigments to phytoplankton groups and then groups to size classes was based on published information regarding phytoplankton pigmentation and morphology (see Table 1 in Vidussi et al. 2001); however, their study was specific to the Mediterranean Sea, and knowledge of phytoplankton size distributions has evolved in the decades since the studies cited in their Table 1. For example, while a few accessory pigments are truly diagnostic as a result of their unique association with a given phytoplankton taxonomic group (e.g., divinyl Chl *a* in prochlorophytes and Allo in cryptophytes), most accessory pigments can be found in multiple phytoplankton groups. Additionally, the assumption that all dinoflagellates and diatoms are microplankton is a known oversimplification, acknowledged by Vidussi et al. (2001) and addressed by subsequent studies for diatoms, but not for dinoflagellates. The reliance of the DPA on links between pigments and phytoplankton taxa (a separate topic addressed with a variety of methods e.g., Mackey et al. 1996; Kramer and Siegel 2019), as well as the size range of different phytoplankton taxonomic groups, is not trivial and requires further investigation.

Numerous works have revised the equations and/or accessory pigment weights used in the DPA, both to tune pigment weights to regional studies or HPLC datasets, and to reduce uncertainties resulting from the inherent assumptions of the DPA (Uitz et al. 2006; Hirata et al. 2008; Uitz et al. 2008; Brewin et al. 2010; Devred et al. 2011; Brewin et al. 2014; Soppa et al. 2014; Brewin et al. 2015, 2017; Di Cicco et al. 2017; Gittings et al. 2019). Many other studies have applied published versions of the DPA equations and pigment weighting coefficients with no revision to values or equations (e.g., Hirata et al. 2011; Brotas et al. 2013; Organelli et al. 2013; Lorenzoni et al. 2015; Zeng et al. 2018). Brewin et al. (2014)

evaluated the DPA using size-fractionated filtration and calculated PSCs using three published variations of the DPA. Their results showed that deviations in PSC estimates among the different DPA models are small compared to the deviations between DPA models and size-fractionated filtration. They also showed that the DPA tends to overestimate nanoplankton and underestimate picoplankton, compared to size-fractionated filtration. However, uncertainties in the size-fractionated filtration data were not quantified, and as a result the authors note that it is difficult to assess which of the two approaches (DPA and size-fractionated filtration) more accurately represents the size structure of the phytoplankton community (Brewin et al. 2014). Previous efforts have also used DPA estimates as the in situ data against which various satellite algorithm results are compared (Brewin et al. 2011), and as the baseline against which satellite-derived particle size distributions were evaluated (Kostadinov et al. 2010). The ongoing application of DPA-based PSCs for describing the phytoplankton community and, perhaps even more importantly, as the in situ standard against which to compare satellite-based PSC algorithms, warrants a closer look at the assumptions and uncertainties of the DPA approach.

Despite the numerous revisions and applications of the DPA, an evaluation of the method using independent measurements of cell size distribution has yet to be conducted. This gap is in large part the result of the labor-intensive task required to quantify cell size for a sufficiently large number of cells and samples. Relatively recent advances in imaging-in-flow cytometry and its application to large-scale ocean studies via continuous flow-through data collection now enable this type of evaluation. Methods of quantifying cell size and size class Chl *a* fractions using cytometry are faced with a different set of challenges from those of the DPA approach to quantifying cell sizes. Cytometry instruments require careful calibration against beads and/or phytoplankton cultures of known size to link scattering intensity or image pixels to cell size. The question of how to define cell sizes – namely by either equivalent spherical diameter, or major axis length – impacts the distribution of cells between size classes. The nature of comprehensively enumerating or capturing images of individual cells means that cytometric

instruments cannot sample large volumes of water, and as a result there are uncertainties in how well a few ml of water sampled using cytometry represent the water mass where discrete samples on the order of 1-2 L of water have been collected for HPLC analysis.

In this study, we use a novel dataset of concurrently measured imaging-in-flow and conventional flow cytometry to evaluate the DPA. We hypothesize that by including information on both cell size and pigments, the DPA can be revised to be a more accurate tool for size class prediction, and uncertainties can be quantified to improve interpretations of using pigments to define size classes. Based on the results of our evaluation, we provide recommendations for future work and for application of uncertainties during implementation of the DPA in prospective studies.

4.2 Materials and Procedures

Data used for evaluation of the DPA were collected as part of the North Atlantic Aerosol and Marine Ecosystems Study (NAAMES). The study consisted of four research cruises in the western North Atlantic Ocean on the R/V *Atlantis* during November 2015, May-June 2016, August-September 2017, and March-April 2018 (Behrenfeld et al. 2019). The study was designed to target different phases of the annual phytoplankton bloom spanning multiple seasons and a wide variety of water types (Fig. 4.1, also see Della Penna and Gaube 2019). Data presented here were collected both while the ship was "on-station" and during transit between stations and to and from port; all transit samples were collected from the ship's continuous flowing seawater system (water intake at approx. 5 m depth), and on-station samples were either collected from the flowing seawater system or from the 5 m depth Niskin bottles that were deployed on a conductivity-temperature-depth (CTD) rosette.

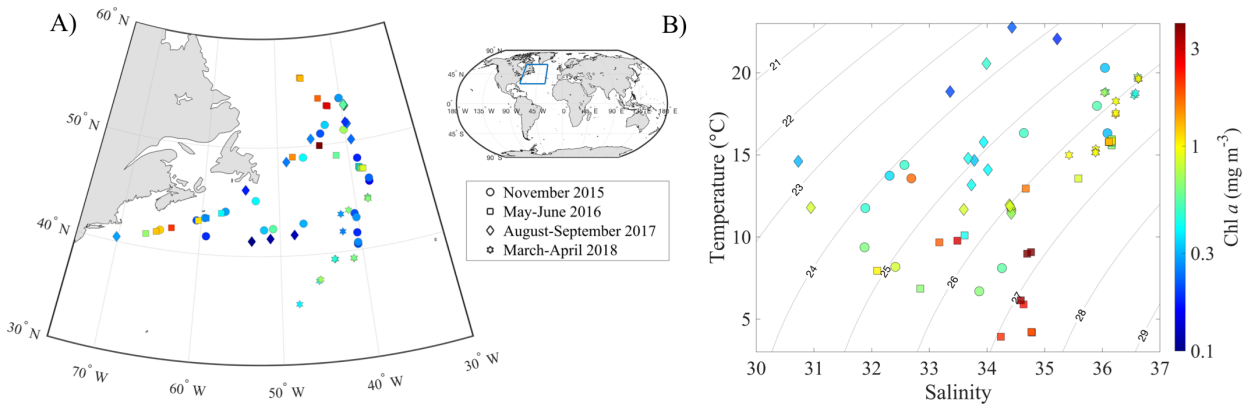

Figure 4.1. Data locations, temperature and salinity values, and Chl *a* concentrations. A) Locations in the western North Atlantic of coincidentally measured cytometry and HPLC data during four research cruises: NAAMES 1 in November 2015 (circles), NAAMES 2 in May-June (squares), NAAMES 3 in August-September 2017 (diamonds), and NAAMES 4 in March-April 2018 (hexagrams) (total n = 90). Samples are colored by Chl *a* concentration. B) Temperature, salinity, and chlorophyll values of the same 90 samples. Values represent averages of +/- 10 min around each sampling location in (A).

4.2.1. HPLC Pigment Data and Application of the DPA

Water samples for HPLC pigment analysis were filtered onto 25 mm diameter, 0.7 μm pore size Whatman GF/F filters and stored in liquid nitrogen until analysis by the Ocean Ecology Laboratory at NASA Goddard Space Flight Center (GSFC) following methods in Van Heukelem and Thomas (2001) and Hooker et al. (2009). Several times per cruise, either duplicates or triplicates were collected by taking independent and consecutive water samples from the flowing seawater system. During HPLC analysis, mean pigment concentrations of duplicates or triplicates were calculated and the standard deviation was used as the uncertainty of HPLC-based Chl *a* in our analysis. Any pigment concentrations below the limit of detection were set to zero. To calculate PSCs from HPLC pigment concentrations, we applied all previously published versions of the DPA (Table 4.1 and references therein). We did not evaluate published relationships for very low chlorophyll concentrations (< 0.08 mg m^{-3}), derived previously for

ultraoligotrphic conditions by Brewin et al. (2010, 2014), as these conditions do not apply to our North Atlantic dataset, where Chl *a* concentrations at our sample locations ranged from 0.18 - 5.14 mg m^{-3}. We evaluated seven unique DPA equations (Table 4.1, Eq. 4.1-4.7), applying the seven diagnostic pigment weights (*W*) reported in Uitz et al. (2006): *W* =[1.41,1.41,1.27,0.35,0.60,1.01,0.86] to the seven pigments: *P*=[Fuco,Peri,Hexa,Buta,Allo,TChlb,Zea]. The same set of weights were applied in Hirata et al. (2008) and Devred et al. (2011). We also tested the impact of using different values for *W* by applying four previously published sets of *W* values for either global or North Atlantic datasets; one set of *W* is reported in each of Brewin et al. (2014); Soppa et al. (2014); Brewin et al. (2015, 2017). Results for picoplankton (*Fp*), nanoplankton (*Fn*), and microplankton (*Fm*) fractions had standard deviations of 0.04, 0.02, and 0.02 (unitless fractions 0-1), respectively, therefore indicating that our results were not significantly changed by the application of different values of pigment weights. Additionally, we computed values of *W* from our dataset using a linear least-squares solver (MATLAB *lsqlin* function), and the summed weighted pigments showed strong correlation (r^2=0.97) with HPLC-derived Chl *a*. We tested the use of *W* values from our dataset in calculating PSCs and our results were not significantly changed. Contributions of each size class to total chlorophyll (*Chl p*, *Chl n*, and *Chl m*) were calculated by multiplying the size class fraction (*Fp, Fn,* or *Fm*) by Chl *a* concentration from HPLC analysis. Following methods of Uitz et al. (2006), we define Chl *a* as the sum of the concentrations of monovinyl Chl *a* + divinyl Chl *a* + chlorophyllide *a* + Chl *a* allomers and epimers.

Table 4.1. Formulas to calculate picoplankton (*Fp*), nanoplankton (*Fn*), and microplankton (*Fm*). Size class fractions are estimated from HPLC phytoplankton pigment concentrations as reported in previously published studies. Devred et al. (2011) and Hirata et al. (2008) apply equations from Uitz et al. (2006) for picoplankton and microplankton, respectively, and so are not shown here to avoid redundancy. *DPw* = summed weighted diagnostic pigments (mg m^{-3}); *DPw* = $\sum_{i=1}^{7} W_i P_i$, where W = pigment weights

(unitless; values shown in text) and P = pigment concentrations (mg m^{-3}) = [Fuco(P_1), Peri(P_2), Hexa(P_3), Buta(P_4), Allo(P_5), TChlb(P_6), Zea(P_7)].

Formula		Reference	Eq.
Picoplankton			
$Fp = \frac{\sum_{i=6}^{7} W_i P_i}{DP_w}$		*Uitz et al. 2006*	(4.1)
$Fp = \frac{W_7 P_7}{DP_w}$	+ all samples with Chl *a* < 0.25 mg m^{-3}	*Hirata et al. 2008*	(4.2)
Nanoplankton			
$Fn = \frac{\sum_{i=3}^{5} W_i P_i}{DP_w}$		*Uitz et al. 2006*	(4.3)
$Fn = \frac{\sum_{i=3}^{6} W_i P_i}{DP_w}$		*Hirata et al. 2008*	(4.4)
$Fn = \frac{(\sum_{i=3}^{5} W_i P_i) + W_1 P_{1,n}}{DP_w}$	where $P_{1,n} = 10^{(0.531 \log_{10} P_3 + 0.708 \log_{10} P_4)}$	*Devred et al. 2011*	(4.5)
Microplankton			
$Fm = \frac{\sum_{i=1}^{2} W_i P_i}{DP_w}$		*Uitz et al. 2006*	(4.6)
$Fm = \frac{(\sum_{i=1}^{2} W_i P_i) - W_1 P_{1,n}}{DP_w}$	where $P_{1,n} = 10^{(0.531 \log_{10} P_3 + 0.708 \log_{10} P_4)}$	*Devred et al. 2011*	(4.7)

4.2.2. Collection and Analysis of Cytometry Data

Size measurements (equivalent spherical diameter) of individual phytoplankton cells within the size range 6-150 µm were obtained from images collected with an Imaging FlowCytobot (IFCB; McLane Research Laboratories, Inc.). The IFCB records images in units of pixels, and size calibration using imaged NIST-traceable size standard beads was used to determine the conversion from pixels to µm. The IFCB

was deployed such that 5 ml of water were drawn automatically from the ship's flow-through system approximately every 20 minutes, and cells and particles were individually imaged if a laser-induced chlorophyll fluorescence threshold was reached (Olson and Sosik 2007). The exact volume of water evaluated for each sample (on average 4.8 ml for our study) was recorded and used during data analysis for quantification of cell concentrations. Images and associated metadata, as well as extracted features (Sosik and Olson 2007), were deposited on EcoTaxa, a web-based platform, developed for the curation and annotation of plankton images (Picheral 2017). Images were classified into taxonomic or functional group categories based on morphological features (images available at https://ecotaxa.obs-vlfr.fr/). Living cells were separated from non-living particles and classification was completed to the highest taxonomic level possible based on instrument resolution and image characteristics. Cells that lack distinct morphological features are difficult to identify from IFCB images, and those cells were grouped under a category of "unidentifiable". A learning set containing 14,917 manually validated images across 84 taxonomic or morphological (e.g., detritus) categories was applied in a random forest machine learning approach to predict the identification of 250,660 images from IFCB samples across the four NAAMES cruises. All images used in this study were then manually confirmed or corrected in their category assignment. Following the export of information on classified particles from EcoTaxa, living cells were grouped into seven broad taxonomic categories: diatoms, dinoflagellates, Dictyochales, prymnesiophytes, cryptophytes, euglenoids, chlorophytes, and "unidentifiable". The only remaining living particles not included in these categories are the relatively rarely imaged ciliates and Rhizaria, which accounted for 2.3% of cell biovolume. Non-living particles including detritus, bubbles, and fecal pellets were removed from the analysis. Cell biovolumes were estimated following Moberg and Sosik (2012), from which equivalent spherical diameter of each cell was calculated. Estimates of cellular carbon for cells > 7 µm were computed following equations in Menden-Deuer and Lessard (2000), and for cells < 7 µm following Sommer et al. (2012).

An Influx Cell Sorter (ICS; BD Biosciences) was used to enumerate cells from discrete water samples collected with 5 m depth Niskin bottles or from the the flow-through system. Samples were pre-filtered through a 64 μm mesh filter prior to analysis. A detailed description of the ISC set-up and operation used during NAAMES field campaigns can be found in Graff and Behrenfeld (2018). To quantitatively assess the size of ICS-measured particles, a calibration was performed using a regression analysis of several phytoplankton cultures against the forward scattering parameter (FSC) recorded by the ICS. The same culture samples were also analyzed by the size-calibrated IFCB, therefore providing sizes of cells from cultures used in the regression. The result is size (equivalent spherical diameter) information for cells recorded by the ICS, ranging approximately from 0.5 μm to 20 μm, which are then used to calculate biovolume and subsequently cellular carbon following Menden-Deuer and Lessard (2000) for cells > 7 μm and Sommer et al. (2012) for cells < 7 μm.

4.2.3. Biovolume Estimates of Pico- Nano- and Microplankton from Merged Cytometry Data

To calculate the micro-, nano-, and picoplankton fractions from the combined IFCB and ICS data (henceforth we refer to the merged data from these two instruments as "cytometry"), we use ICS data for picoplankton (< 2 μm), IFCB data for microplankton (20 – 150 μm), and a combination of the two data types for nanoplankton (2 – 20 μm). To determine the size threshold for defining the contribution of the two instruments to the nanoplankton we analyzed the total particle size distribution of IFCB samples. As a result of reduced counting efficiency near the lower size limit of the IFCB, we identify the point at which the particle size distribution begins to decrease as the minimum size at which the IFCB provides quantitative information on particle concentrations. This minimum size threshold changes with instrument gain settings, and potentially also with the physiology of phytoplankton populations. We calculate and apply size thresholds based on threshold averages calculated for each cruise, using 6.3, 7.9, 8.8, and 7.1 μm for the four chronological NAAMES cruises, respectively. Nanoplankton were thus

estimated for each sample using the combined ICS data below these size thresholds and IFCB data above these size thresholds for each cruise.

IFCB and ICS samples were considered to be co-located if the time difference between the two did not exceed three hours. Larger phytoplankton are less abundant for a given sample volume and therefore there is a higher uncertainty in their count compared to small cells. To address this bias, and to increase statistical power for counts of large cells, if the nearest matching IFCB sample had fewer than 2,500 images, additional IFCB samples were added until the 2,500 image threshold from within a +/- 3 hour time span is reached. To avoid combining in-line IFCB samples that were collected as the ship passed through different water masses, we removed any samples that deviated from the co-located IFCB/ICS sample by more than 36% in terms of the number of IFCB images per sample. This 36% threshold was determined by one standard deviation above the mean (84th percentile) of the coefficient of variation of cell counts per sample for all samples within a maximum of +/- 3 hours of the nearest matching ICS and IFCB sample in time. This approach is motivated by the fact that, although there is some natural variability in cell counts within a given water mass, any change in cell counts outside the 84th percentile likely indicates the transition of the ship into a new water mass. This criterion resulted in the elimination of 75 out of 450 initial IFCB samples (17%) for a total of 375 remaining IFCB samples. The end result of matching IFCB samples with coincident ICS samples is a dataset of 90 IFCB/ICS matches with cellular biovolume and carbon estimates across all three phytoplankton size classes. Finally, IFCB/ICS pairs were matched with corresponding HPLC samples. Nearly all HPLC and ICS water samples were collected simultaneously; of the samples that do not match exactly in time, all HPLC samples were within three hours of the nearest ICS sample. The total cell biovolume of these data were compared to HPLC-determined Chl *a* concentrations to assess how well the cytometry data represent the phytoplankton community at a given time and place (Fig. 4.2A). Cytometry cell carbon was converted to units of Chl *a* by applying the mean carbon-to-chlorophyll (C:Chl) values derived from two sources: 1) the model of

Behrenfeld et al. (2016) applied to the NAAMES data by Fox et al. (2020), and 2) inline measurements of particulate backscattering ($b_{bp}(\lambda)$) converted to phytoplankton carbon following Graff et al. (2015) and divided by Chl *a* concentrations, where Chl *a* is derived from particulate absorption ($a_p(\lambda)$) red peak line height as in Boss et al. (2013). Refer to the NASA SeaBASS data repository for $a_p(\lambda)$ and $b_{bp}(\lambda)$ data (https://seabass.gsfc.nasa.gov/). Despite the best guess values of C:Chl from both modeled and in situ data, there is still uncertainty in the accuracy of the conversion from carbon to Chl *a*. To show this uncertainty, we use minimum and maximum known values of C:Chl from the literature (Sathyendranath et al. 2009) and calculate the range of Chl *a* converted from cell carbon to display as uncertainties in Fig. 2B.

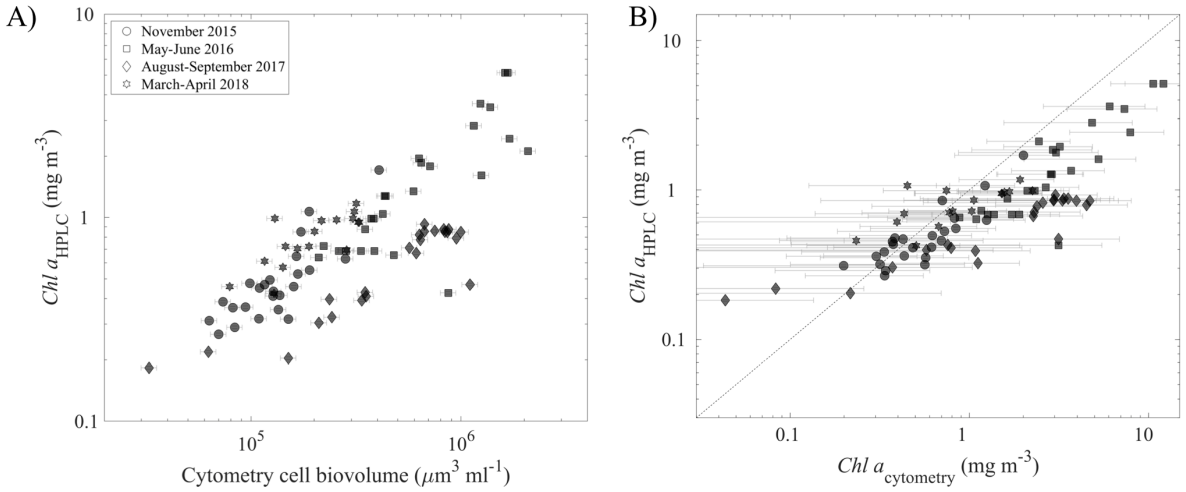

Figure 4.2. Cell biovolume and Chl *a* compared to Chl *a* from HPLC. A) Total cell biovolume from cytometry and Chl *a* concentration from HPLC across all 90 match-up samples. Y-axis error bars are the uncertainty in Chl *a* from HPLC based on replicate samples. X-axis error bars are 17% uncertainty in IFCB biovolume as estimated by Moberg and Sosik (2012). In both panels, marker shape represents the different NAAMES cruises. B) Chl *a* estimated from cytometry data following conversion of cell biovolume to cell carbon, and carbon to Chl. Y-axis error bars as in A), and x-axis error bars depict the range of Chl *a* values possible based on published C:Chl values.

4.2.4. Statistics Used for Method Evaluation

Contributions of the three phytoplankton size fractions to Chl *a* as derived from the DPA and cytometry approaches were compared using the Pearson correlation coefficient (r^2; calculated using MATLAB *corr* function), and four additional metrics, defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i^D - X_i^C\right)^2} \tag{4.8}$$

$$rRMSE = 100\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i^D - X_i^C}{0.5(X_i^D - X_i^C)}\right)^2}\right) \tag{4.9}$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^D - X_i^C\right) \tag{4.10}$$

$$rBias = 100\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i^D - X_i^C}{0.5(X_i^D - X_i^C)}\right)\right) \tag{4.11}$$

where *n* is the number of samples, *X* is a given PSC (either the unitless fraction or the Chl *a* concentration), and the superscripts *D* and *C* denote the PSCs determined by either the DPA or from cytometry, respectively. As both the pigment and cytometry-based PSC estimates have uncertainties and outliers, we used a mean of the two for calculating the relative RMSE (rRMSE) and relative bias (rBias).

4.3. Assessment and Discussion

4.3.1. Overall Trends in PSC Estimate Comparisons

Cell biovolume from cytometry and Chl *a* from HPLC show a linearly increasing trend for log-normalized data across all four cruises (Fig. 4.2A). Variability around this relationship is expected as the result of variation in Chl *a* per cell volume that occurs among different phytoplankton taxa as well as growth conditions, e.g., nutrients and light. Application of estimated C:Chl values results in a Pearson correlation of $r^2 = 0.79$ between Chl *a* concentrations from HPLC and cytometry (Fig. 4.2B). The overall overestimation of Chl *a* by cytometry is mostly driven by NAAMES 2 (Fig. 4.2B square symbols) and NAAMES 3 (Fig. 4.2B diamond symbols), and could be the result of phytoplankton populations with different species compositions and/or growth conditions, which in turn have different C:Chl values than are represented by the mean values applied.

The two DPA equations to estimate the Chl *a* fraction of picoplankton (Table 4.1, Eq. 4.1 and 4.2) show different trends when compared to cytometry (Fig. 4.3A): Eq. (4.1) significantly overestimates the relative contribution of picoplankton to Chl *a*, and Eq. (4.2) significantly underestimates the picoplankton fraction (Table 4.2). We found poor correlation between the nanoplankton fraction calculated from all three DPA equations (Table 4.1, Eq. 4.3-4.5) and the nanoplankton fraction from cytometry (Fig. 4.3B; Table 4.2). For the majority of the match-ups, DPA estimates of the nanoplankton fraction were lower than those from cytometry. For the microplankton, Eq. 4.6 and 4.7 (Table 4.1) both overestimate the microplankton fraction and show low correlations with cytometry (Fig. 4.3C; Table 4.2). Chl *a* concentrations associated with each size class show the same bias patterns as the PSC fractions (Fig. 4D-F; Table 4.3). In most cases, correlations are higher when considering Chl *a* concentrations of the three size classes, and in some cases the values of relative bias (reflecting over- or underestimation) are improved when considering Chl *a* concentrations of PSCs (Table 4.3).

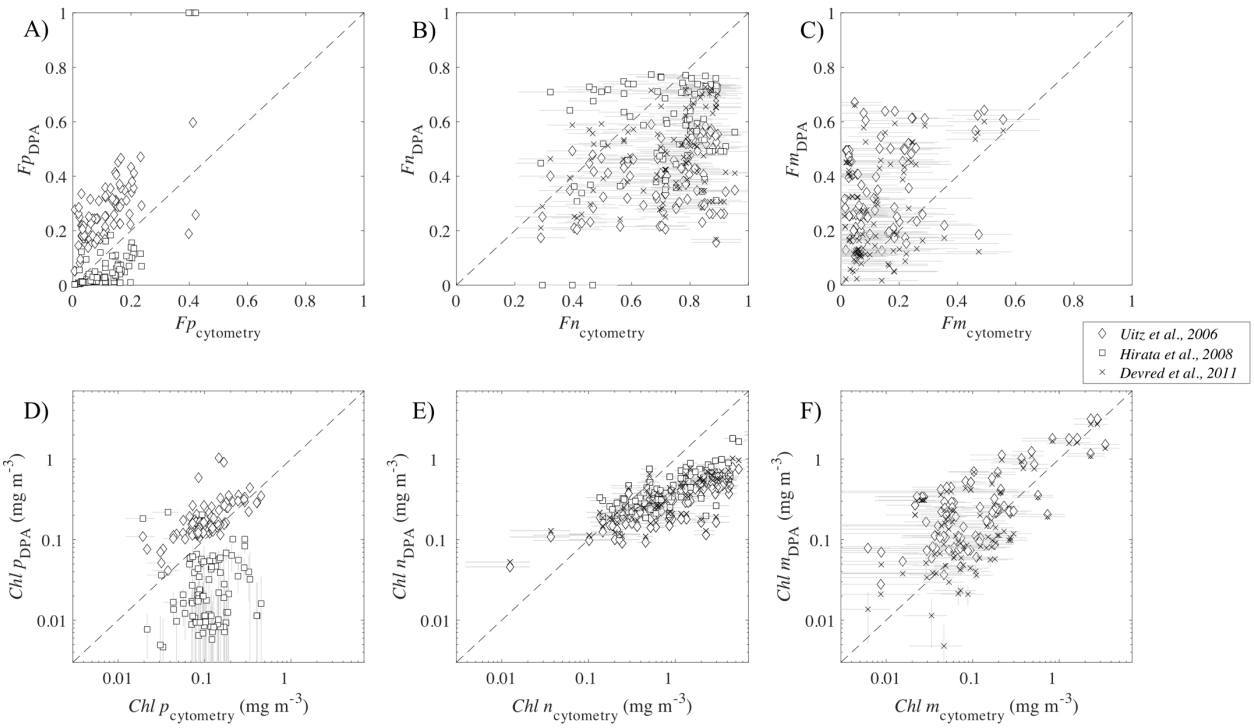

Figure 4.3. Comparison of size classes estimated by cytometry and pigments. Fractions of A) pico-, B) nano-, and C) microplankton calculated using cytometry (x-axes) and previously published DPA equations (y-axes; see Table 4.1 for equations). Error bars in the x-axis direction are determined by statistical counting error and uncertainty in estimation of cell biovolume. D-F) Chl *a* concentrations of each size class calculated from the DPA (y-axes) and cytometry (x-axes). Error bars in the y-axis direction represent the uncertainty in Chl *a* concentrations determined from replicate samples. X-axis error bars represent uncertainty in the conversion from C:Chl based on the published values of C:Chl for each size class.

Table 4.2. Statistics of PSC fractions estimated using published versions of the DPA compared to cytometry data.

	Eq.	Reference	r^2	RMSE (unitless)	rRMSE (%)	Bias (unitless)	rBias (%)
Fp	(4.1)	*Uitz et al. 2006*	0.58	0.15	94	0.13	79
	(4.2)	*Hirata et al. 2008*	0.73	0.13	114	-0.05	-96
Fn	(4.3)	*Uitz et al. 2006*	0.34	0.38	69	-0.34	-72
	(4.4)	*Hirata et al. 2008*	0.35	0.26	38	-0.15	-20
	(4.5)	*Devred et al. 2011*	0.35	0.32	54	-0.25	-72
Fm	(4.6)	*Uitz et al. 2006*	0.31	0.28	116	0.21	96
	(4.7)	*Devred et al. 2011*	0.34	0.23	102	0.12	61

Table 4.3. Statistics of PSC Chl *a* concentrations estimated using published versions of the DPA compared to cytometry data.

	Eq.	Reference	r^{2*}	RMSE (unitless)	rRMSE (%)	Bias (unitless)	rBias (%)
Chl p	(4.1) x Chl *a*	*Uitz et al. 2006*	0.67	0.15	60	0.07	44
	(4.2) x Chl *a*	*Hirata et al. 2008*	0.09	0.13	136	-0.10	-119
Chl n	(4.3) x Chl *a*	*Uitz et al. 2006*	0.76	1.81	112	-1.19	-101
	(4.4) x Chl *a*	*Hirata et al. 2008*	0.75	1.59	90	-0.94	-62
	(4.5) x Chl *a*	*Devred et al. 2011*	0.75	1.65	100	-1.04	-81
Chl m	(4.6) x Chl *a*	*Uitz et al. 2006*	0.68	0.37	97	0.13	65
	(4.7) x Chl *a*	*Devred et al. 2011*	0.62	0.35	95	0.05	27

*calculation of r^2 performed on the log transformed Chl *a* data

4.3.2. Discrepancies Between Attribution of Pigments to PSCs, and Observed Cell Sizes and Groups

A summary of the analysis on the discrepancies between pigment and size class assignments is shown in Table 4.4. Note that the lower size limit of the IFCB (approx. 6-7 µm) allows for the identification of phytoplankton groups within the nano- and microplankton size classes, but not the nanoplankton between 2-6 µm or the picoplankton size class. Table 4.4 also makes reference to cyanobacteria, which are not imaged by the IFCB (in their single-celled form; *Trichodesmium* colonies can be imaged), but are identified by the ICS.

Table 4.4. Assignments of the seven accessory pigments used in the DPA to phytoplankton groups and size classes. Open boxes with dashed outlines indicate where groups are known to fall within the picoplankton size range but are not observed by the IFCB, which only quantifies cells > approx. 7 μm.

Diagnostic pigment	Abbreviation	Phytoplankton group(s)	Pico*	Nano*	Micro
Fucoxanthin	Fuco	Diatoms, Prymnesiophytes, Dictyochales			
Peridinin	Peri	Dinoflagellates			
19'-Hexanoyloxy-fucoxanthin	Hexa	Prymnesiophytes			
19'-Butanoyloxy-fucoxanthin	Buta	Prymnesiophytes, Dictyochales			
Alloxanthin	Allo	Cryptophytes			
Total Chlorophyll *b*	TChl*b*	Euglenoids, Chlorophytes			
Zeaxanthin	Zea	Cyanobacteria †			

 = DPA-based pigment size class assignments

 = Cell sizes of phytoplankton groups observed with IFCB imagery

*Cyanobacteria, picoeukaryotes, and nanoplankton < 7 μm are not measured by the IFCB. These groups were sampled with the ICS; however, individual cell size of phytoplankton groups estimated by manual gating of ICS data were not analyzed in this study.

†Note that some cyanobacteria colonies, such as that of *Trichodesmium*, can reach microplankton size or larger.

Picoplankton are defined by Uitz et al. (2006) as the weighted sum of TChlb and Zea, and by Hirata et al. (2008) as the weighted value of Zea alone plus all samples with Chl *a* concentrations < 0.25 mg m^{-3} (Table 4.1, Eq. 4.1 and 4.2, respectively). The result of this approach is an overestimation of picoplankton when all TChlb is assigned to the picoplankton size class, and an underestimation when Zea alone is assigned to picoplankton (Fig. 4.3A,D). Zea is found primarily in cyanobacteria, which usually exist as picoplankton. TChlb is found in eukaryotic picoplankton, such as representative members of the phylum Chlorophyta (and more specifically some prasinophytes) (Roy et al. 2011). However, TChlb-containing phytoplankton groups, namely euglenoids and chlorophytes, belonging to both the nanoplankton and microplankton size classes were identified with IFCB imagery (Fig. 4.4). Note that "unidentifiable" bars in Fig. 4.4 may also include TChlb-containing organisms. Thus it is not surprising that when all TChlb is assigned to picoplankton (Eq. 4.2), the DPA will overestimate the relative contribution of picoplankton. Other groups (that contain neither Zea or TChlb) but could be contributors to the picoplankton community include prymnesiophytes (Moon-Van Der Staay et al. 2000) and Bolidophyceae (Guillou et al. 1999; Kuwata et al. 2018).

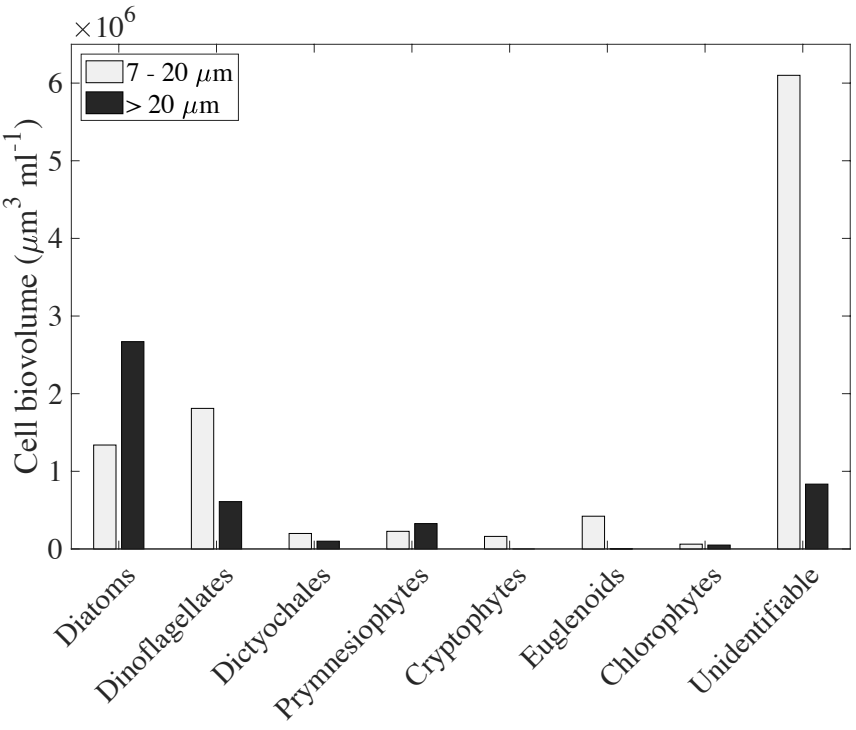

Figure 4.4. Summed biovolume of phytoplankton cells imaged by the IFCB across all samples. Note that many dinoflagellates as well as some prymnesiophytes and cryptophytes have been identified, but because they are often more difficult to ID with high certainty, some cells of these types were classified as "unidentifiable" and hence biovolumes of these groups are likely underestimated. The "unidentifiable" category may also include other nanoeukaryote groups such as chlorophytes and pelagophytes.

The presence of nanoplankton is defined by the DPA as either the weighted sum of Allo, Buta, and Hexa (Table 4.1, Eq. 4.3); the same combination of pigments plus TChlb (Table 4.1, Eq. 4.4); or as in Eq. 4.3 but with some proportion of the Fuco reallocated from microplankton to nanoplankton (Table 4.1, Eq. 4.5). Allo is a diagnostic pigment for cryptophytes, which we observed in the nanoplankton size class in IFCB imagery (Fig. 4.4). Buta is found in prymnesiophytes such as *Phaeocystis* and coccolithophores, which are known to span all three size classes (Jeffrey et al. 1997; Roy et al. 2011) and within our IFCB imagery data are observed both within nano- and microplankton size classes (Fig. 4.4). Hexa is also found in

prymnesiophytes, as well as in silicoflagellates (e.g., Dictyochales) and pelagophytes. Dictyochales are identifiable in IFCB imagery and fall within both the nano- and microplankton size classes (Fig. 4.4). TChlb, assigned by Eq. 4.4 to nanoplankton, is found in chlorophytes and euglenoids; however, among chlorophytes, prasinophytes are known as important members of the picoplankton (Roy et al. 2011), and we also observed some microplankton-sized chlorophytes (Fig. 4.4). The relative contribution of diatoms to the nanoplankton has historically been overlooked, but recently received more attention (Leblanc et al. 2018). Our imagery data show that 33% of all diatom biovolume for cells > 7 µm falls within the nanoplankton size class. This result is likely an underestimation of nanoplankton diatom biovolume, due both to the possibility of unidentified cells in the IFCB imagery, and/or the presence of diatoms 2-7 µm in size. We also note that chain-forming taxa such as *Chaetoceros* are evaluated for cell size by the length of the chain, not the individual cells. Prymnesiophytes and Dictyochales also both span the nano- and microplankton size range and contain Fuco. Devred et al. (2011) assume Hexa and Buta are only present in the nanoplankton size class, and the relationship between Hexa, Buta, and Fuco is then used to attribute some proportion of Fuco to nanoplankton rather than microplankton. Applying the adjustment to the DPA by Devred et al. (2011) to account for the Fuco present within nanoplankton attributes, on average, 44% (+/- std of 29%) of Fuco in our dataset to nanoplankton. Microplankton are overestimated by the DPA (Table 4.1, Eq. 4.6 and 4.7), both when defined as the weighted sum of all Fuco and Peri (Eq. 4.6), and when defined by the weighted sum of all Peri and the remaining fraction of Fuco not attributed to nanoplankton (Eq. 4.7; Fig. 4.3C, F). Of identifiable dinoflagellates, we observe 74% within the 7-20 µm nanoplankton size class (Fig. 4.4). This result is notable in its indication that current DPA equations are incorrectly assigning Peri, which is found exclusively in dinoflagellates, to the microplankton (Table 4.1, Eq. 4.6 and 4.7).

Overall, identification of nano- and microplankton cells using IFCB imagery reveals the extent to which several major phytoplankton groups are misrepresented in the DPA framework. Although previous work

on development of the DPA has often acknowledged its assumptions and potential sources of uncertainty, we now have the cytometry and phytoplankton group data needed to address how these assumptions could impart biases to estimated PSCs. Most notably, the attribution of all Peri (found in dinoflagellates) to microplankton by all published DPA microplankton equations, as well as the attribution of TChlb to picoplankton in the most commonly applied version of the DPA from Uitz et al. (2006), results in inaccurate PSC fractions and PSC Chl *a* concentrations.

4.4. Recommendations

As a result of the direct comparisons between DPA and cytometry presented in this study, we recommend a new set of DPA equations for estimating phytoplankton size classes. Based on the presence of accessory pigments in major phytoplankton groups (Roy et al. 2011; Jeffrey et al. 1997) and the size ranges of phytoplankton groups both from the literature and as observed in this study, several adjustments are made to the assumptions of current DPA methods. While our data are regional, the assignment of pigments to phytoplankton size classes is not specific to groups found within the North Atlantic. The original assignment of diagnostic pigments to phytoplankton groups and groups to size classes made by Vidussi et al. (2001) and cited by the DPA versions evaluated here was developed for a study in the Mediterranean Sea and has subsequently been applied in numerous global studies. We therefore recommend an update to the original pigment-group-size class assignments based on our open ocean dataset as an incremental step forward in improving the accuracy of estimating PSCs from HPLC pigments.

We define the fraction of picoplankton as the weighted sum of Zea and half of the TChlb to account for the presence of TChlb in both pico- and nanoplankton (Eq. 4.12). Although some TChlb- containing groups can be microplankton, including some euglenoids, *Halosphaera*, and *Pterosperma*, we observed these infrequently, and so distribute the TChlb between pico- and nanoplankton, which are known to

contain common marine genera including *Micromonas*, *Pryramimonas*, and other prasinophytes, as well as *Euglena* and *Pterosperma* (Tomas 1997). The nanoplankton fraction is defined by the weighted sum of Hexa, Buta, and Allo, the remaining half of TChlb, half of the Fuco concentration, and 75% of the total Peri concentration (Eq. 4.13). The assignment of Hexa, Buta, and Allo to nanoplankton are unchanged from previously published DPA equations. Microplankton-sized prymnesiophytes and silicoflagellates that may contain Hexa and Buta were present in our samples; however, the potential for prymnesiophytes in both the unidentifiable nanoplankton category and in the picoplankton renders it difficult to know how to accurately redistribute Hexa or Buta out of nanoplankton. The microplankton fraction is defined as the weighted sum of the remaining half of the Fuco combined with the remaining 25% of Peri (Eq. 4.14).

$$Fp = \frac{0.5W_6P_6 + W_7P_7}{DP_w} \tag{4.12}$$

$$Fn = \frac{0.5W_1P_1 + 0.75W_2P_2 + \sum_{i=3}^{5} W_iP_i + 0.5W_6P_6}{DP_w} \tag{4.13}$$

$$Fm = \frac{0.5W_1P_1 + 0.25W_2P_2}{DP_w} \tag{4.14}$$

We evaluate Eq. 4.12-4.14 using the pigment weighting values (*W*) from Uitz et al. (2006) for consistency in evaluation against performance metrics of Eq. 4.1-4.7. We also tested values of *W* derived from our North Atlantic dataset and the results were not significantly different. Results show improved relationships for all three size classes, both in relative fractions and Chl *a* concentrations (Fig. 4.5; Table 4.5). In particular, the values of absolute and relative bias are reduced in all cases, which is an improvement even when overall uncertainty due to natural variability cannot be eliminated. Despite reallocated pigments and adjusted DPA equations, there is a clear fractionation of phytoplankton into size classes by the DPA, whereas the cytometry data displays a continuum. For example, Figure 4.5B

shows the Chl *a* fraction of nanoplankton mainly falling within the range of 60-80%, whereas the Chl *a* fraction of nanoplankton from cytometry spans 20-100%.

The simplified proportions of pigments allocated to each size class by the updated DPA method presented here cannot represent the complexity that underlies the allocation of pigments to phytoplankton groups and size classes in the ocean. We have shown that with several simple adjustments, some inaccuracies of current DPA methods are reduced. However, it is important to consider how natural variability may impact the interpretation of results derived using the DPA approach. Results from the cytometry data highlight the fact that phytoplankton groups containing different pigments – and within different size classes – are present in the ocean at different times, and this leads us to a broader question of whether the definitions of the pico-, nano-, and microplankton size classes are necessarily appropriate.

In addition to recommending the use of Eq. 4.12-4.14 to estimate PSCs, we also encourage the application of uncertainties when calculating estimated PSC fractions or Chl *a* concentrations. This manuscript provides quantified uncertainties associated with previously published DPA equations (Tables 4.2 and 4.3), as well as the uncertainties for the updated methods presented here (Eq. 4.12-4.14; Table 4.5). With the rapid advancement of technology for quantifiable measurements of plankton (Lombard et al. 2019), we recommend further similar studies to ours that independently evaluate the size structure of phytoplankton communities, so that as the DPA is applied for the study of ocean ecosystems, for satellite algorithm development, and during ocean modeling efforts, its inherent biases and uncertainties are appropriately considered.

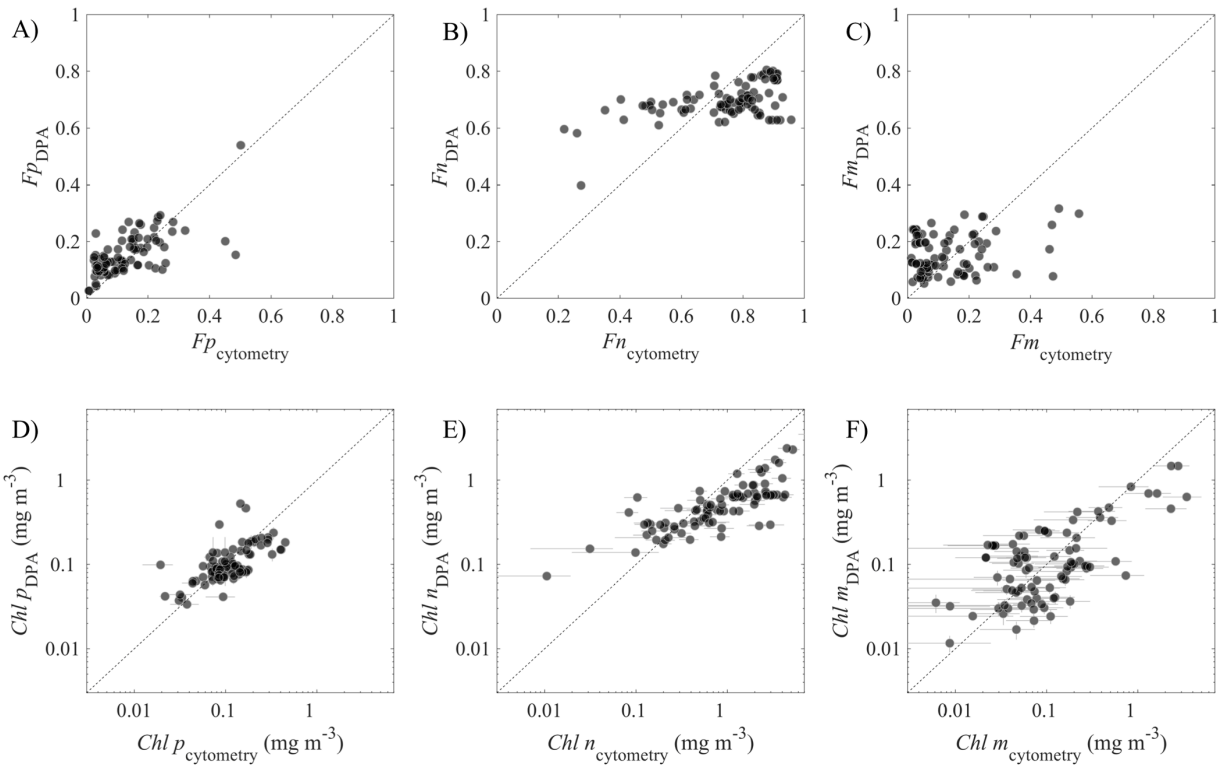

Figure 4.5. Comparison of size classes estimated by cytometry and pigments using updated equations. A-C) Fractions of pico-, nano-, and microplankton, respectively, calculated using cytometry (x-axis) and Eq. 4.12-4.14 (y-axis). Error bars in the x-axis direction are determined by statistical counting error and uncertainty in estimation of cell biovolume. D-F) Chl *a* concentrations of each size class calculated from Eq. 4.12-4.14 (y-axes) x Chl *a* , and cytometry (x-axes). Error bars in the y-axis direction represent the uncertainty in Chl *a* concentrations determined from replicate samples. X-axis error bars represent uncertainty in the conversion from cell carbon to Chl *a* (C:Chl) based on the published values of C:Chl for each size class.

Table 4.5. Statistics of PSC fractions and Chl *a* concentrations estimated using updated DPA equations compared to cytometry data.

	Eq.	r^2	RMSE (unitless)	rRMSE (%)	Bias (unitless)	rBias (%)
Fp	(4.12)	0.70	0.07	63	0.03	38
Fn	(4.13)	0.51	0.13	19	-0.06	-6
Fm	(4.14)	0.20	0.12	86	0.03	38
Chl p	(4.12) x Chl *a*	0.66*	0.09	44	-0.01	-7
Chl n	(4.13) x Chl *a*	0.84*	1.40	78	-0.87	-49
Chl m	(4.14) x Chl *a*	0.65*	0.42	84	-0.10	-3

*calculation of r^2 for Chl *a* concentrations performed on the log transformed Chl *a* data

CHAPTER 5

DIATOM CARBON FROM SPACE: A MODEL FOR THE WESTERN NORTH ATLANTIC OCEAN

5.1 Introduction

Variability in phytoplankton community composition (PCC) both spatially and through time directly affects transfer of carbon to higher trophic levels (Dickman et al. 2008; Brett et al. 2000), patterns in carbon export (Siegel et al. 2014), and biogeochemical cycles (Dutkiewicz et al. 2013). Studies show that phytoplankton communities are changing on ocean basin scales due to shifts in environmental conditions (Hays et al. 2005; Rousseaux and Gregg 2015; Neukermans et al. 2018), and the use of satellite-based information is critical for continued efforts to observe and understand changes in ocean ecosystems. Diatoms, the focus of the present study, are estimated to contribute between 35 and 75% of total primary production in the ocean, depending on the region (Nelson et al. 1995). Their aggregation and sinking characteristics can result in sequestration of carbon to the deep ocean (Honjo and Manganini 1993) and they are the preferred food of copepods (Irigoien et al. 2002), which are the major food source for many planktivorous fish. Diatoms are generally assumed to be favored by high turbulence and nutrient availability (Margalef 1978; Wyatt 2014), although the diversity of diatoms and their deviation from the high-turbulence-nutrient dogma has received increased attention in recent years (Glibert 2016; Leblanc et al. 2012). Although they are extremely diverse with an estimated 200,000 different species (Armbrust 2009), diatoms are often considered in the literature as a uniform functional group and as exclusively microplankton (> 20 μm).

Assessing patterns in diatom populations at ocean-basin scales is vastly improved by satellite-based measurements. Efforts to estimate diatoms from space have included analysis of multispectral water leaving radiance anomalies (Alvain et al. 2005; 2008), remote-sensing reflectance ($R_{rs}(\lambda)$) band ratios (Sathyendranath et al. 2004), and a method of Differential Optical Absorption Spectroscopy (DOAS) that

requires high spectral resolution reflectance measurements in the blue end of the visible spectrum (Bracher et al. 2009; Sadeghi et al. 2012). In the majority of past studies, diatoms – whether their abundance, dominance, or biomass – were estimated both during model development and evaluation via pigment proxies. However, defining diatoms on the basis of diagnostic pigments and pigment ratios faces several significant challenges. First, accessory pigments are rarely unique to a single taxonomic group, even when groups are defined broadly. The ratio of fucoxanthin to Chl *a* has been applied as a diagnostic pigment for diatoms (Hirata et al. 2011; Soppa et al. 2014). However, fucoxanthin is found in several common phytoplankton groups in addition to diatoms, namely prymnesiophytes and silicoflagellates (Jeffrey and Vesk 1997; Roy 2011; Table 1.2). Changes in pigment ratios driven by variability in nutrients and photosynthetically available light are also a source of uncertainty when defining phytoplankton groups by relative pigment concentrations (Goericke and Montoya 1998). Additionally, converting pigment concentrations to cell biomass or carbon content is non-trivial, as the ratio of pigments to cell carbon is a function of growth conditions and phytoplankton types, and therefore highly variable both between and within phytoplankton groups (Sathyendranath et al. 2009). The model presented by Raitsos et al. (2008) uses a neural network approach with various satellite-based input parameters to estimate phytoplankton groups, including diatoms, from space. However, the model relies on data from the Continuous Plankton Recorded (CPR), which does not provide quantitative estimates of diatom abundances but rather their "probability of occurrence". A second study using a neural network approach to estimate the fractional contribution of four phytoplankton groups, including diatoms, to Chl *a* relies on pigment-based phytoplankton group definitions and on the output of the NASA Ocean Biogeochemical Model (NOBM, Gregg et al. 2003), and is thus limited both by defining PCC based solely on pigments, and in spatial resolution by the NOBM 1° x 1° grid (Palacz et al. 2013).

Hirata et al. (2011), hereafter denoted “H11”, estimates contributions of diatoms to total Chl *a* by first defining diatoms as a function of fucoxanthin concentration, and then deriving an equation to estimate

diatoms as a function of Chl *a*. While this approach can be considered a reasonable first-order approximation of diatom Chl *a*, it does not account for the variability in the relationship between diatoms and fucoxanthin, and in turn the relationship between diatoms and Chl *a*. With the exception of divinyl chlorophylls and zeaxanthin, all phytoplankton accessory pigments in the ocean covary both with each other and with Chl *a* (Trees et al. 2000; Pan et al. 2010; Kramer et al. 2019), which has been shown to result in Chl *a*-based algorithms with comparable uncertainties to more complex optical algorithms (e.g., Chase et al. 2017). It is therefore necessary to identify times, places, and environmental conditions when phytoplankton populations deviate from what can be estimated from known covariability between Chl *a*, accessory pigments, and major phytoplankton groups. Alvain et al. (2005; 2008) accounted for the covariability between Chl *a* and phytoplankton groups in their remote-sensing-based model by using water-leaving radiance values that in turn were normalized to an average Chl *a*-based spectrum (determined from a large database of matching Chl *a* and water-leaving radiance satellite measurements). However, their assumption is that relative diatom concentration is based on the value of fucoxanthin concentration relative to Chl *a* concentration, which ultimately constrains the accuracy of the model given the limitations with using pigments to define PCC outlined above. We note that analysis of phytoplankton pigments is extremely valuable in many applications of understanding phytoplankton communities (e.g., Swan et al. 2016; Kramer et al. 2018). However, when used as a proxy for defining phytoplankton groups without quantification of uncertainties, pigments should be carefully considered.

Two key elements are required for robust satellite-based PCC algorithm development: first, it is necessary to account for the covariation of Chl *a* with other pigments and phytoplankton groups. Second, appropriate uncertainty propagation that includes best estimates of uncertainties associated with both model inputs and targets, as well as error statistics of algorithm performance. Imaging-in-flow cytometry data for assessment of phytoplankton groups provides an opportunity to address the first of these key elements, and in combination with uncertainty propagation, enables the development of an

algorithm to estimate diatom carbon from space. In the present study we apply quantitative cell imagery to address variability in diatom carbon biomass in relation to Chl *a* concentrations, by investigating correlations between environmental parameters and binned values of Chl *a* concentration. We apply this knowledge in combination with data on environmental conditions to train a shallow neural network to produce estimates of diatom carbon from both in situ and satellite-based data. The concept of applying a neural network to the estimation of phytoplankton groups is not novel (Raitsos et al. 2008; Palacz et al. 2013). However, the imagery-based target data used in the model presented here is novel, and its use leads to a neural network that accounts for the variability in diatom carbon as a function of Chl *a*. The model we present here for the western North Atlantic region is the first to provide estimates of diatom carbon from space that are based on in situ information from quantitative cell imagery.

5.2 Study Region and Datasets

The North Atlantic Ocean is a highly productive region of global ocean and the annual phytoplankton accumulation there accounts for approximately one fifth of net global CO_2 uptake (Sabine et al. 2004). PCC in the region is highly variable both seasonally and regionally (Barton et al. 2015). The North Atlantic Aerosol and Marine Ecosystems Study (NAAMES) was conducted during 2015-2018 in the western North Atlantic, during different seasons and stages of the phytoplankton annual cycle (Behrenfeld et al. 2019). Temperature and salinity in surface waters (~5 m depth) were determined in situ from continuous shipboard measurements by an SBE9+ thermosalinigraph. Estimates of diatom carbon were derived from imagining-in-flow cytometry and HPLC pigment analysis as described below.

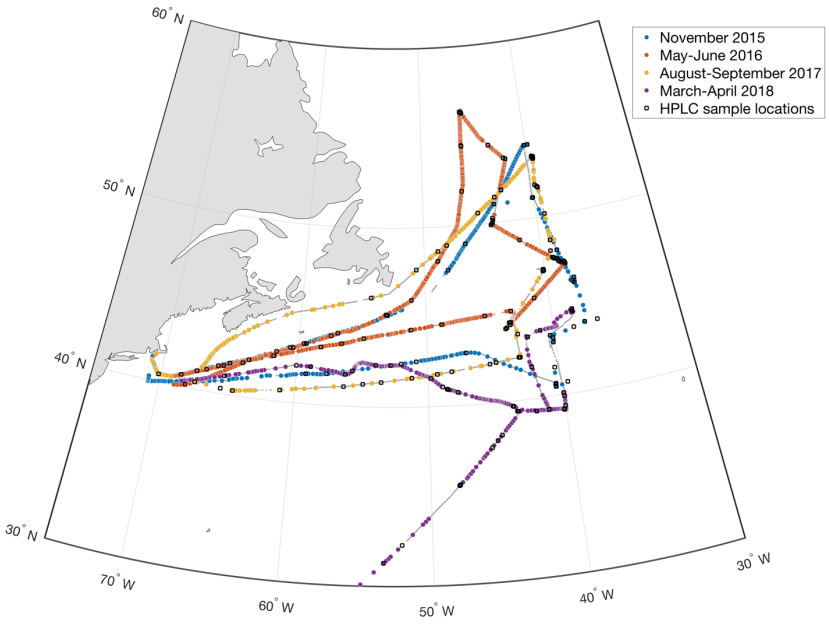

Figure 5.1. Locations of the four NAAMES cruise tracks in the western North Atlantic Ocean. Colored dots show IFCB sample locations following the procedure to combine samples, n = 1,755. Black squares show the location of discrete water samples taken for HPLC analysis, n = 205.

5.2.1 Imaging-in-flow Cytometry

Morphological information on individual phytoplankton cells was collected using an Imaging FlowCytobot (IFCB, MacLean Research Laboratories, Inc.). The IFCB provides quantitative information on cell concentrations in the size range of approximately 6 - 150 µm (Olson and Sosik 2007). The IFCB was connected to the flowing seawater system of the research vessel (R/V Atlantis), where it automatically pulled 5 ml (nominally) of water approximately every 20 minutes. Within the IFCB, particles are channeled single-file and imaged with a camera if a laser-induced chlorophyll fluorescence threshold just upstream of the camera is reached. Quantitative information on particle size and morphology is calculated for each image, including maximum and minimum axes lengths, area, biovolume, and equivalent spherical diameter (ESD) derived from particle volume estimates (Sosik and Olson 2007; Moberg and Sosik 2012; https://github.com/hsosik/ifcb-analysis).

5.2.2 Plankton Classification and Carbon Estimates

We used a combination of manual and machine learning approaches to classify IFCB images into morphological and taxonomic categories. All data were uploaded to EcoTaxa, a web platform for both viewing and classifying images (Picheral et al. 2017). All images used in this study are available for viewing at https://ecotaxa.obs-vlfr.fr/ (search "NAAMES"). Initially, a learning set was built by manually sorting 14,700 images into 58 categories (Table A.1). The learning set was applied within the EcoTaxa platform using a random forest, machine learning approach to categorize 4.8 million IFCB images from the four NAAMES cruises. Of these images, 2.2 million were either manually confirmed or corrected in their categorization. These images were in turn labeled at two taxonomic resolutions: the highest level possible based on morphological features (usually genus level), and a broader categorization of 18 groups; the latter results in an increased number of cells per category (Table A.2). The 18 categories were used in a deep learning classification model that predicts the category of all classes with 86% accuracy, and diatoms as a group with 90% accuracy (see Appendix). This resulted in a collection of 336,872 diatom cells imaged and identified from the entire NAAMES investigation. Biovolume, ESD, and concentrations were determined on a per ml basis by summing all cells for a given sample and normalizing by the volume sampled, enabling comparison of diatom populations across all IFCB samples. Diatom cell biovolume was calculated using Moberg and Sosik (2012), and subsequently cellular carbon was calculated following the diatom-specific formula reported by (Menden-Deuer and Lessard 2000). Diatom carbon concentrations across all data samples are log-normally distributed, and as a result the log-transformed carbon concentration values are used during model development described below in section 5.3.1.

To reduce statistical counting errors associated with larger cells occurring at lower concentrations, IFCB samples were combined when total particles imaged per sample is fewer than 2,500. Cell information

from IFCB samples was combined until the total particle count threshold of 2,500 cells is reached or the combined samples exceeds a ±3 hours. To ensure that combined samples were not drawn from different water masses we removed samples that had a total particle count deviating from the first in the set of combined samples by more than one standard deviation (±36%). The result of combining samples to decrease statistical counting errors reduces the number of data points available for analysis to 1,755 out of the original 4,733 IFCB samples (Figs. 5.1, A.1-A.2). Combined samples with fewer than 20 diatoms were further removed, and the total number of data points used in the analysis was 1,449. See Fig. A.3 for the distribution of the number of diatoms in each of the 1,499 samples. The 95% confidence interval value for the upper limit of relative uncertainty in a cell count of 20 is 50%, and the relative uncertainty increases as the number of cells counted decreases (Clopper and Pearson 1934). This motivates our removal of data points with fewer than 20 diatoms, as the particle count uncertainties are high enough that the subsequent calculations of carbon concentrations are highly uncertain.

5.2.3 HPLC Pigments and CHEMTAX

Concentrations of phytoplankton accessory pigments were determined from discrete surface water samples using High Performance Liquid Chromatography (HPLC) (n = 205, Fig. 5.1). Water samples were collected either from the ship's flowing seawater system with intake at ~5 m, or from the 5 m Niskin bottle deployed on a CTD rosette. Water (1-2 L) was filtered immediately onto 25 mm diameter GF/F filters (0.7 µm nominal pore size) and the filters were stored in liquid nitrogen for the duration of the cruise, shipped on dry ice and processed by the Ocean Ecology Laboratory at NASA Goddard Space Flight Center following Van Heukelem and Thomas (2001) and Hooker et al. (2009). Two different approaches to estimate relative and absolute diatom chlorophyll from pigments were employed. The first is the Diagnostic Pigment Analysis (DPA), which assigns pigments to phytoplankton groups (Claustre 1994; Vidussi et al. 2001; Uitz et al. 2006; H11). We note that the application of diagnostic pigments was

originally used to define phytoplankton size classes (pico-, nano-, and microplankton) (Vidussi et al. 2001; Uitz et al. 2006), but has subsequently been used to define phytoplankton taxonomic groups (H11; Soppa et al. 2014). The fraction of total Chl *a* attributed to diatoms is defined by the DPA as:

$$Diat_{\mathrm{DPA}} = \frac{1.41\mathrm{Fuco}}{\sum \mathrm{DP_w}} \tag{5.1}$$

where summed weighted diagnostic pigments are defined as ΣDP_w = 1.41Fuco + 1.41Peri + 1.27Hexa + 0.6Allo + 0.35Buta + 1.01Chlb +0.86Zea and are highly correlated with Chl *a* and represent the major phytoplankton groups (Uitz et al. 2006). Pigments abbreviations are as follows: Fuco = fucoxanthin, Peri = peridinin, Hexa = 19'-hexanoyloxyfucoxanthin, Allo = alloxanthin, Buta = 19'-butanoyloxyfucoxanthin, Chlb = total chlorophyll b, and Zea = zeaxanthin. We tested the use of four other sets of published weighting coefficients (Brewin et al. 2014; Soppa et al. 2014; Brewin et al. 2015, 2017) and our results were not significantly changed. The Chl *a* concentration of diatoms is then calculated by:

$$DiatChl_{DPA} = Diat_{DPA} * (Chl\ a) \tag{5.2}$$

which has units of mg m^{-3}. The fraction of Chl *a* attributed to diatoms was also calculated using the CHEMTAX program (Mackey et al. 1996). CHEMTAX requires a set of phytoplankton groups and associated initial ratios of accessory pigments to Chl *a* to be defined, thereby attributing accessory pigments to phytoplankton groups (and allowing for the same pigment to be present in multiple groups). We used the initial pigment ratios and phytoplankton groups from two different published sources: Swan et al. (2016), and van de Poll et al. (2013) (see Appendix; Figs. A.4-A.8). The output from CHEMTAX is the fraction that each phytoplankton group contributes to Chl *a*, which in turn is multiplied by Chl *a* to get the Chl *a* concentration of each group, including diatoms.

Comparison of imagery-derived and pigment-derived diatoms in the same units requires the conversion of either imagery to units of Chl *a*, or pigments to units of carbon. Here we choose to convert the diatom Chl *a* estimated from pigment concentrations ($DiatChl_{DPA}$) to units of carbon, for comparison with cell imagery measurements. This conversion is non-trivial as there is large variability in the ratio of these two cell quantities, commonly referred to as the carbon-to-chlorophyll (C:Chl) ratio. Previous studies have assessed C:Chl variability as a function of season and region, phytoplankton group, and photoacclimation (Behrenfeld et al. 2016; Sathyendranath et al. 2009; Jackson et al. 2017). To address the known variability in C:Chl, we use a mean value of C:Chl for diatoms suggested by Sathyendranath et al. (2009), as well as the minimum and maximum values from the same study to calculate and apply a range of uncertainty values to diatom carbon estimates from pigment-based methods:

$$DiatCarb_{DPA\_}mean = 41 * DiatChl_{DPA} \quad (5.3)$$

$$DiatCarb_{DPA\_}max = 107 * DiatChl_{DPA} \quad (5.4)$$

$$DiatCarb_{DPA\_}min = 15 * DiatChl_{DPA} \quad (5.5)$$

$$DiatCarb_{DPA\_}range = DiatCarb_{DPA\_}\max - DiatCarb_{DPA\_}min \quad (5.6)$$

This range of values is also within the range of C:Chl estimated for the NAAMES cruises by both modeling and using in situ optical estimates of phytoplankton carbon and Chl *a*, which average 28 ± 11, 27 ± 10, 94 ± 43, and 27 ± 11 mg m^{-3}:mg m^{-3} for the four NAAMES cruises (Fox et al. 2020). The only exception is the subtropical region of the NAAMES03 cruise (sampled during 2-10 September 2017), which had high C:Chl values ranging between 70 and 235. As a result of low values of both carbon and Chl *a* during this time period (mean Chl *a* concentration of 0.17 mg m^{-3}), there are higher uncertainties in C:Chl values; however, also as a result of low biomass, data from this time also do not significantly impact our estimates of diatom carbon.

5.2.4 Pigments Estimated from Particulate Absorption Spectra ($a_p(\lambda)$)

Hyperspectral $a_p(\lambda)$ were measured using an ac-s instrument (SeaBird Scientific) with water pumped continuously from the ship's flowing seawater system, using a calibration independent method and binned to 1-minute temporal resolution (Slade et al. 2010). Concentrations of five pigments/pigment groups and associated uncertainties were then estimated from the spectra following Chase et al. (2013): Chl *a*, Chl *b*, Chl *c*, and grouped photosynthetic and photoprotective carotenoids (PSC and PPC, respectively). As previously mentioned, phytoplankton accessory pigments covary with Chl *a* and each other in the ocean. To account for this and obtain information independent from the covariation with Chl *a*, we determined a pigment anomaly:

$$\Delta Pigment = Pigment_{a_p} - Pigment_{global} \quad (5.7)$$

where $Pigment_{ap}$ is one of Chl *b*, Chl *c*, PSC, or PPC derived from $a_p(\lambda)$, and $Pigment_{global}$ is the value of a given pigment expected from global pigment covariation relationships, as defined in Chase et al. (2017). The Δ*Pigment* values therefore provide information on whether a given accessory pigment is anomalously high or low relative to what is expected based on global pigment covariations.

5.2.5 Satellite Data

Three satellite-based products are used: sea surface temperature (SST), sea surface salinity (SSS), and Chl *a*. The level 3 Chl *a* product, calculated using band ratio algorithms from multispectral reflectance data (Hu et al. 2012), was downloaded from the NASA OC.DAAC for NASA's MODIS Aqua instrument (https://oceancolor.gsfc.nasa.gov/l3/). The global monthly composite data for May 2016 was used, corresponding with the field campaign of the second NAAMES cruise. SMAP ocean surface salinity for May 2016 was downloaded from the NASA PO.DAAC (https://podaac.jpl.nasa.gov/), and the MUR SST, a

merged, multi-sensor product, was also downloaded from the PO.DAAC for May 11, 2016. Satellite data are subset to the western North Atlantic region defined between the latitudes 30-60°N and longitudes 75-30°W. Data are linearly interpolated without extrapolation onto the latitude/longitude SST grid, which is the finest spatial resolution of the three data types (Chl *a*, SSS, and SST). This results in Chl *a*, SSS, and SST values at 0.1 degree resolution for a 4501 x 3001 sized grid (Fig. 5.2).

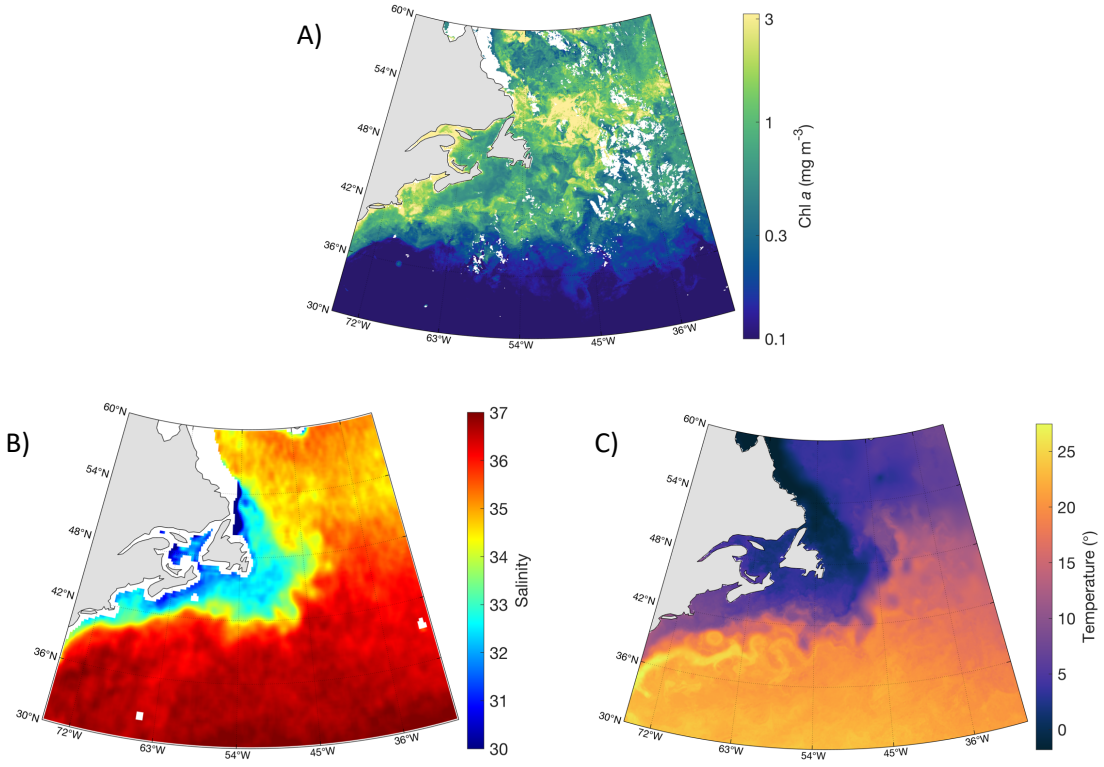

Figure 5.2. Satellite data used for neural network model input. A) MODIS Aqua L3 Chl *a* concentration monthly composite for May 2016. B) SMAP ocean surface salinity, monthly composite for May 2016. C) SST from the MUR merged, multi-sensor L4 product for May 11, 2016.

5.3 Approach

We used the following sequence of analyses in the present study, with the methods of step 3 described in detail in sections 5.3.1 and 5.3.2:

1. Compare diatom carbon estimated from the pigment-based and imagery-based approaches.

2. Investigate relationships between environmental parameters and variable diatom carbon for a given Chl *a* concentration.

3. Train a shallow neural network to estimate diatom carbon using environmental parameters that are significantly correlated with imagery-based diatom carbon.

4. Run the neural network with satellite inputs to demonstrate the applicability of the neural network framework to estimate diatom carbon from space.

5.3.1 Diatom carbon neural network setup and testing

Neural networks allow for interactions between variables that are more complex than regression models between individual variables and model targets. Here we focus on environmental input parameters that both can be obtained from remote sensing data, and have known relationships to diatoms (Brun et al. 2015). The values of both the input parameters and diatom carbon span a wide range across the study region (Figs. 5.3, A.9), and therefore capture the variability in the relationships between the chosen environmental parameters and diatom biomass (carbon) across different seasons and in different water masses throughout the western North Atlantic.

A neural network was trained, validated, and tested using the MATLAB *nftool* function with five input parameters: Chl *a*, salinity, temperature, day length in hours, and day of year. All five parameters are either positively or negatively significantly correlated with diatom carbon calculated from IFCB imagery (Figs. A.9, A.10). The parameters serve as proxies for environmental conditions that influence the presence of diatoms. For example, temperature and salinity serve as proxies for water masses that may be more or less favorable to diatom growth. Day length and day of the year serve as proxies for stratification intensity and date of stratification or destratification onset, as well as the time of maximal

zooplankton grazing. The measurements of input parameters are simultaneous to the target diatom carbon measurements, and so the water mass history, which strongly influences the phytoplankton population at a given time and place, is not accounted for. However, the coincident environmental and diatom measurements are related via the assumption of environmental proxies, and are what is available given the nature of the underway ship measurements used in this study. Individual input parameters are not sufficient to provide predictive power for diatoms with associated uncertainties that are acceptably low. However, their combination in the context of a neural network approach reduces uncertainties in predicting diatom carbon. Patterns between input and target parameters that are non-linear and/or have critical values (Fig. A.10) can be captured by a neural network. The network was trained using a process of Bayesian regularization, where updates to weight and bias values are made according to Levenberg-Marquardt optimization (MATLAB Deep Learning Toolbox™ Reference, 2020). The network is a two-layer feedforward network, with a hidden layer that uses a sigmoid transfer function, and an output layer that uses a linear transfer function. Through iterative testing, we found that a network with 11 neurons in the hidden layer performed the best in terms of median error between the modeled and target diatom carbon. The input data are an array of 5 x 1449 values; the five parameters listed above coinciding with the 1,449 target values of diatom carbon for all IFCB samples used in the analysis. The inputs and target were randomly separated into independent training (n = 1,232) and testing (n = 217) datasets. Network training continued until a minimum performance gradient was reached, i.e., the validation errors failed to decrease. This occurred with a run time of several seconds and therefore thresholds for maximum epochs and training time were not a limiting factor. Once the network was trained using in situ input and target data, and function for the network was saved and run using satellite data as the inputs (Fig. 5.4).

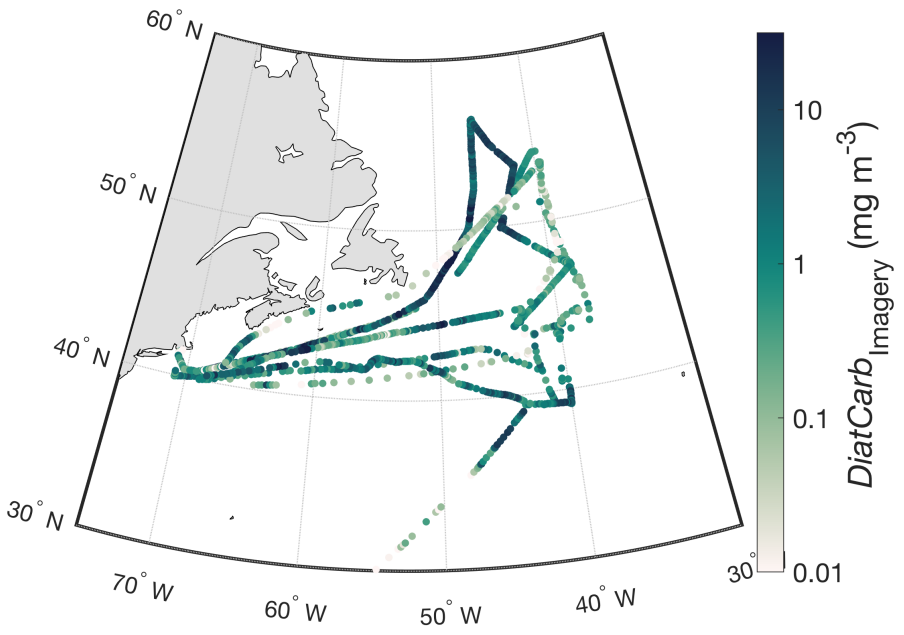

Figure 5.3. Diatom carbon (mg m^{-3}) estimated from cell imagery; n = 1,449.

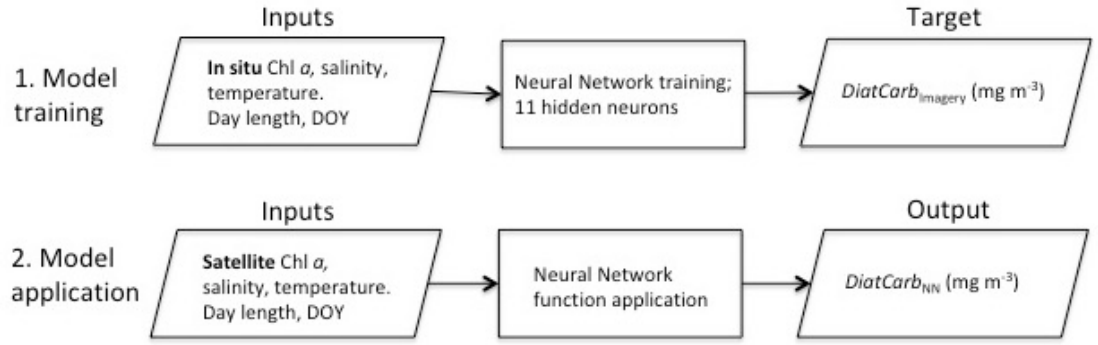

Figure 5.4. Flow chart of neural network model training and application.

5.3.2 Evaluating Model Uncertainties

Uncertainties must be accounted for both in the data collected for model inputs and targets, as well as in the accuracy of the network model itself. In situ Chl *a* concentrations determined via Gaussian decomposition of $a_p(\lambda)$ (see section 5.2.4) have a relative root-mean-square error (RMSE) value of 29%

when compared to HPLC Chl *a* concentrations (Fig. A.11). In our network model, input parameter uncertainties are assumed to be zero for values of day length, day of year, and negligible in situ-measured water temperature and salinity. Uncertainty in diatom carbon from imagery is determined by combining the error in estimating biovolume from an image, which is reported in Moberg and Sosik (2012) as 17% (the same uncertainty applies to updated versions of the biovolume method available at https://github.com/hsosik/ifcb-analysis, H. Sosik *pers. comm.*), with the uncertainty resulting from statistical counting errors, which was determined following Eq. (ii) in Clopper and Pearson (1934) and as applied in Gaube et al. (2017) and has a mean value of 18%. Also combined to these uncertainties is the 10% error associated with the accuracy of identifying all diatoms using the image classification network (see Appendix). All errors are relative (percentage) values, allowing for combination together following:

$$DiatCarb_{Imagery}\_Unc = \sqrt{0.17^2 + 0.18^2 + 0.1^2 + 0.29^2}, \quad (5.8)$$

which has a value of 0.39 (39%), assuming the uncertainties are uncorrelated. The neural network used to estimate diatom carbon has uncertainty associated with the accuracy of the model itself (Figs. 5.5 and 5.6). Following the model training, errors are calculated by the difference between the model target ($DiatCarb_{\text{Imagery}}$) and the model output ($DiatCarb_{\text{NN}}$). The median value of the error in estimated diatom carbon is 1.03 mg m^{-3} and the relative error is 52% (Fig. 5.6). This relative error is combined with the uncertainty in diatom carbon shown above:

$$DiatCarb_{NN}\_Unc = \sqrt{0.39^2 + 0.52^2}, \quad (5.9)$$

resulting in a value of 0.65, or 65%. Note that at low estimated diatom carbon values, the absolute error will dominate over the relative error.

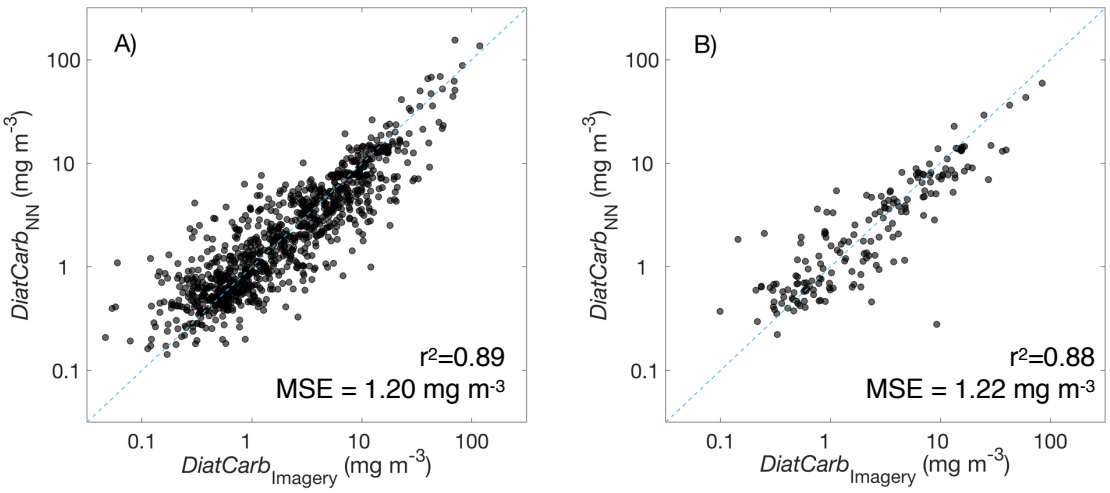

Figure 5.5. Results of neural network training and testing. A) Network training results, n = 1,232. B) Network testing results, n = 217; data points are random and independent of the data used for training. r^2= Pearson correlation coefficient; MSE = median square error.

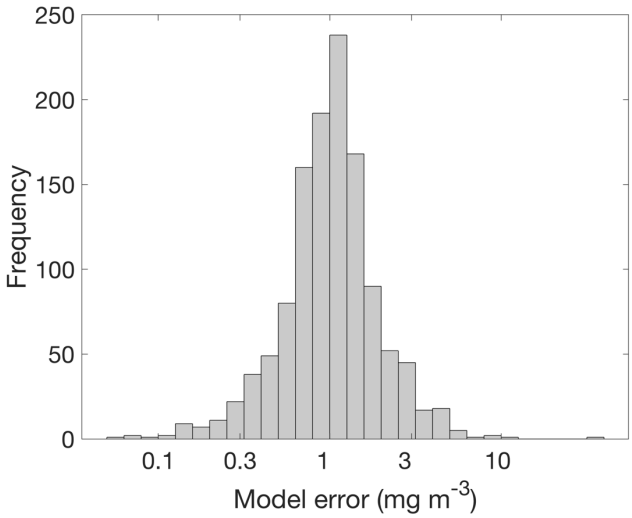

Figure 5.6. Frequency distribution of neural network model error. Error is defined as the difference between diatom carbon from imagery (model target), and the diatom carbon values predicted from the model. Median error in estimates of diatom carbon = 1.03 mg m^{-3}; median relative error = 52%; n = 1,449.

5.4 Results and Discussion

5.4.1 Diatoms Estimated from HPLC Pigments

Diatom carbon estimated from HPLC pigments following the DPA approach shows significantly higher values when compared to those obtained from cell imagery, with relative RMSE and bias values of 146% and 119%, respectively (Fig. 5.7). This is likely explained by the assumption in the DPA method that all the fucoxanthin in a sample is associated with diatoms. It is well known that other groups of phytoplankton such as prymnesiophytes, silicoflagellates, and pelagophytes contain fucoxanthin (Jeffrey et al. 1997; Roy et al. 2011) and therefore attribution of all fucoxanthin to diatoms is likely to overestimate their concentration. CHEMTAX analysis also shows overestimation of diatom carbon compared to imagery-based results. The degree of overestimation varies, depending on the initial pigment ratio and phytoplankton group inputs used in the CHEMTAX model (Figs. A.3-A.7). There are likely cells smaller than the lower detection limit of the IFCB (~6 μm) that contain fucoxanthin, in particular prymnesiophytes (e.g., coccolithophores) and pelagophytes. Some small nanoplankton and picoeukaryotes that contain fucoxanthin could potentially be diatoms. However, previous studies suggest that only an estimated few percent of diatoms occur as small nanoplanktonic types in the open ocean (Leblanc et al. 2012; Leblanc et al. 2018).

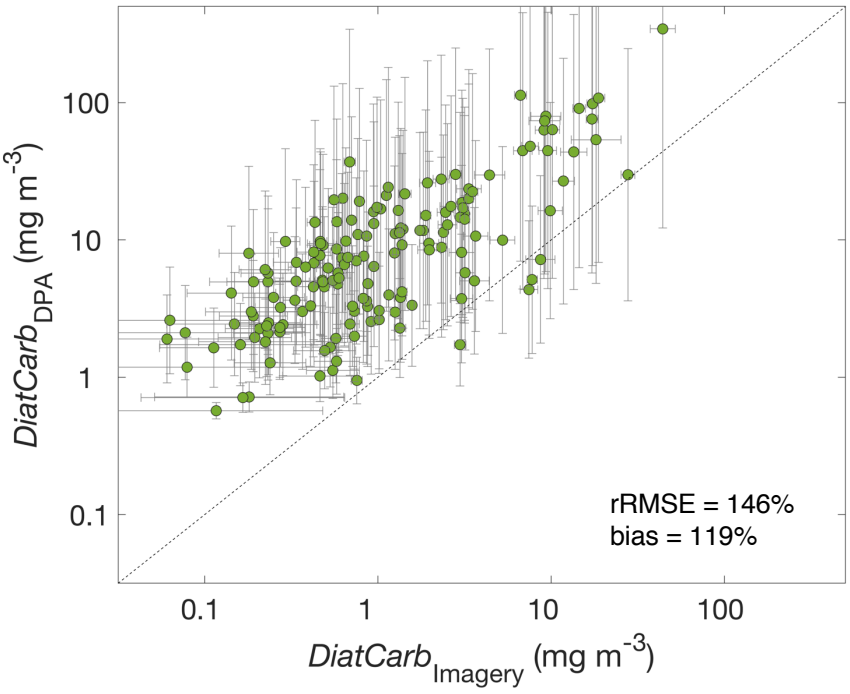

Figure 5.7. Diatom carbon estimated from IFCB imagery and from the Diagnostic Pigment Analysis (DPA) following H11. Error bars on the x-axis represent the uncertainty in cell biovolume estimates combined with statistical counting errors. Error bars on the y-axis represent the range of possible values when converting from DPA-based diatom Chl *a* values and carbon (Eq. 5.6). Relative RMSE (rRMSE) and percent bias are calculated by normalizing by the mean values of the two datasets; neither imagery or DPA-based diatom carbon is assumed to be "truth".

5.4.2 Diatom Variability as a Function of Chl *a*

The algorithm of H11 provides estimates of the contribution of diatoms to Chl *a* following:

$$DiatChl_{H11} = ([1.3272 + \exp(-3.9828x + 0.1953)]^{-1}) * \text{Chl } a, \quad (5.10)$$

where x = $\log_{10}$(Chl *a*). The results of applying the H11 algorithm (or subsequent similar algorithms, e.g. that of Soppa et al. 2014) do not capture the variability in diatoms as a function of Chl *a*. In contrast, this

variability is detected by imagery data (Fig. 5.8), and is important as it affects the predictive capability of Chl *a*. It is also not accounted for when defining diatoms as a function of fucoxanthin as Chl *a* and fucoxanthin are tightly correlated (Fig. A.12). The H11 model performs well when diatoms are a greater proportion of the total carbon of all living particles imaged by the IFCB, and while limited in capability to describe diatom variability as a function of Chl *a*, the H11 model can potentially provide an upper bound for diatom Chl or carbon (Fig. 5.8).

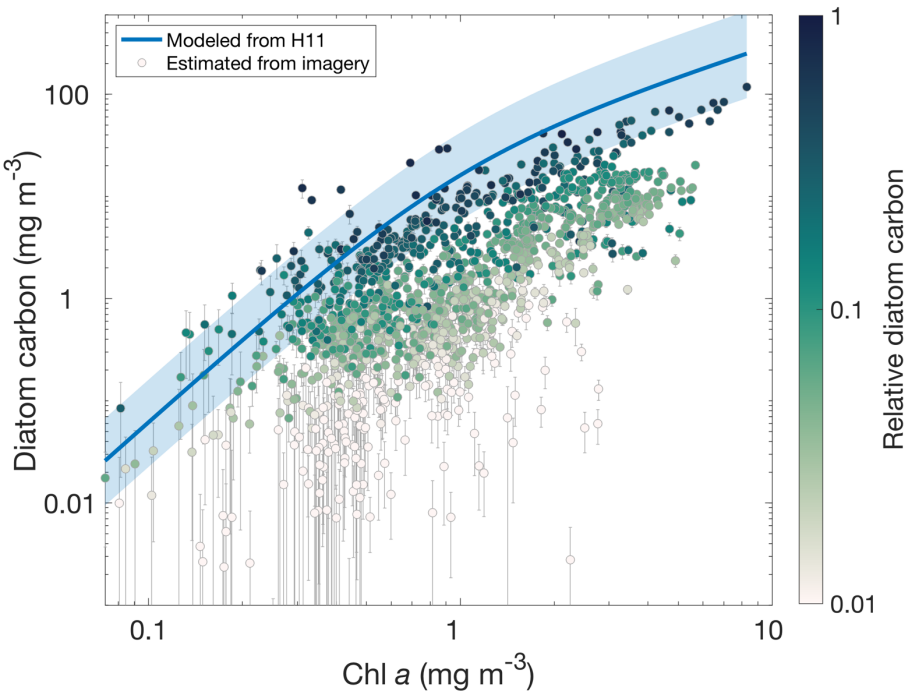

Figure 5.8. Diatom carbon estimated from IFCB imagery and using the formula of H11 compared to Chl *a*. Diatom carbon values represent the concentration for 1,449 IFCB samples. Data points from imagery are colored by the diatom carbon relative to total particle carbon from IFCB measurements, and error bars represent the combined uncertainty in particle biovolume estimates and statistical counting errors. Light blue filled area around the diatom carbon estimated using the H11 formula represents the range of values in diatom carbon with variable C:Chl conversions.

To investigate what determines the variability in diatoms carbon at different Chl *a* concentrations, we analyzed correlation coefficients (non-parametric Spearman's ρ) between 8 environmental parameters and four log-normally distributed Chl *a* concentration bins (0.1-0.3, 0.3-1, 1-3, 3-10 mg m^{-3}) (Fig. 5.9). Relative diatom carbon is the most strongly correlated, with ρ values > 0.6 for all four Chl *a* concentration bins. However, as relative diatom carbon is not an independent measurement from absolute diatom carbon and requires knowledge of PCC carbon for its calculation, it is not a useful parameter for predicting diatom carbon. At high Chl *a* concentrations (3-10 mg m^{-3}), accessory pigment anomalies are negatively correlated with diatom carbon from imagery. The majority of remaining correlation coefficients are not significant, indicating that there are not predictive relationships between individual parameters and diatom carbon, for any given bin of Chl *a* concentration. This finding motivates the use of a neural network, which has the capability of modeling the complex, non-linear relationships between diatoms and environmental parameters that cannot otherwise be modeled using linear or polynomial equations.

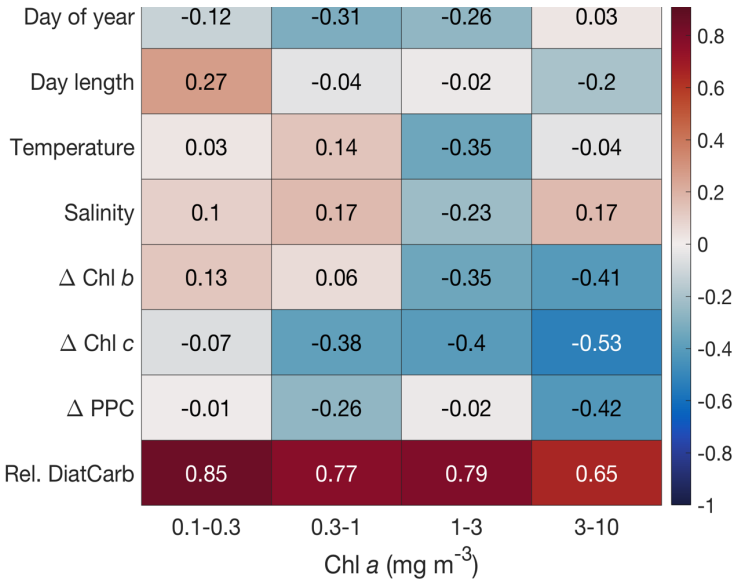

Figure 5.9. Correlations between diatom carbon and environmental parameters across four Chl *a* concentrations. Correlations are Spearman's rank correlation (ρ) values, Chl *a* values are logarithmically

spaced concentration bins (0.1-0.3, 0.3-1, 1-3, 3-10 mg m^{-3}), and there are eight parameters of environmental conditions, pigment anomalies, and relative diatom carbon (the latter denoted Rel. DiatCarb). The positive correlations between relative diatom carbon and Chl *a* across all concentration bins can also be seen in the colored dots of Fig. 5.8.

5.4.3 Satellite-based Diatom Carbon

The same network model that is trained with the five input parameters: Chl *a*, SSS, SST, day length, and day of year, is run with Chl *a*, SSS, and SST from satellite. The output is diatom carbon in units of mg m^{-3} ($DiatCarb_{NN}$) at the spatial resolution of the input satellite data (Fig. 5.10A). Clear spatial patterns emerge with increased diatom carbon associated with some coastal regions, as well as in waters near the northern boundary of the Gulf Stream. The Gulf Stream also stands out as a distinct band of low-diatom water, which is expected based on the low levels of diatoms associated with the warmer, more oligotrophic water delivered to the region by the Gulf Stream. Regions of the map that are outside the ranges of the input parameters should be considered to potentially have higher uncertainties, as the accuracy of the neural network extrapolating to conditions that were not used originally to train the model cannot be quantified. Applying the method of H11 to estimate diatom carbon as a function of Chl *a* ($DiatCarb_{H11}$) shows patterns that are similar in some ways to the results of the network output, but in general are much higher and closely follow, as expected, the spatial patterns of Chl *a* (Fig. 5.10B). The difference of the two models is defined as:

$$\Delta DiatCarb = DiatCarb_{H11} - DiatCarb_{NN}, \quad (5.11)$$

and shows overestimation of diatom carbon by the DPA relative to imagery for most parts of the observed region, and especially in more northern latitudes (Fig. 5.10C, 5.11). Model uncertainty calculations (section 5.3.2) show a mean value of 65%.

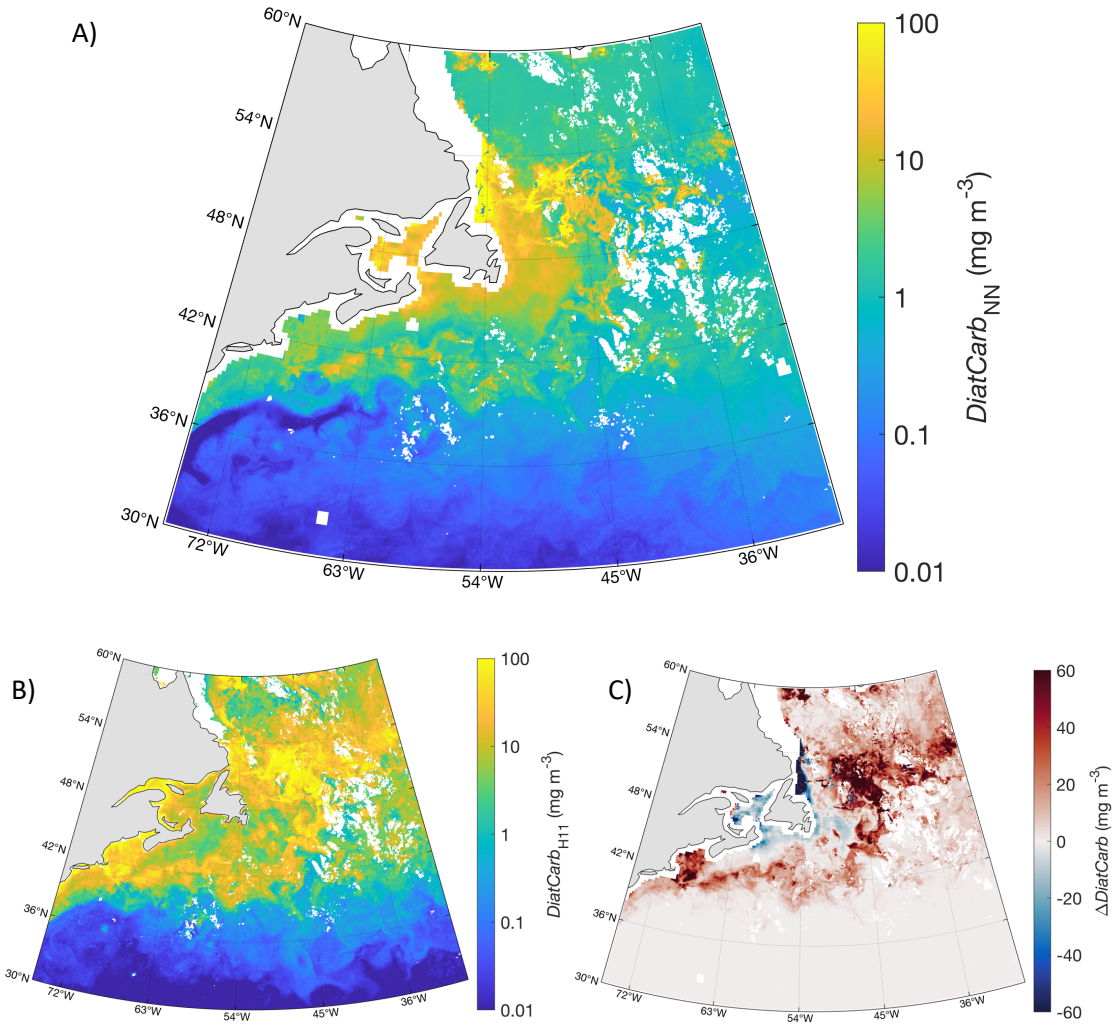

Figure 5.10. Diatom carbon estimated from satellite data for the northwest Atlantic Ocean in May 2016. A) Diatom carbon estimated with the 5-parameter neural network model described in this study (*DiatCarb*$_{NN}$). B) Diatom carbon estimated using the formula of H11 (*DiatCarb*$_{H11}$). C) The difference between the H11 model and the network model of the present study (Δ*DiatCarb*).

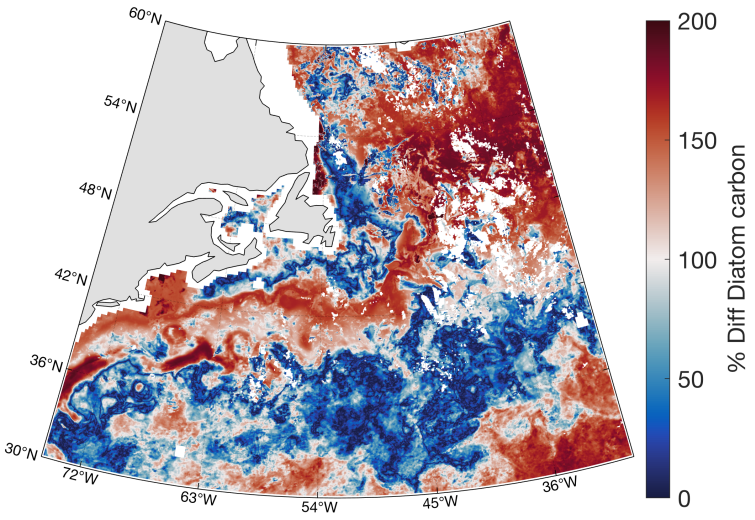

Figure 5.11. Percent difference in diatom carbon. Values are calculated as the absolute difference between the H11 model ($DiatCarb_{H11}$) and the neural network model ($DiatCarb_{NN}$), normalized to the mean of the two.

For model results using satellite data input, additional uncertainty will be imposed by both the accuracy to which the satellite data estimates in situ parameters, as well as the environmental variability within a given pixel. The extent to which the value of a satellite pixel parameter (e.g., a given Chl *a* concentration or salinity value) represents the waters within the pixel size (ranging from 0.1 to 0.25 degrees square for the data used in the present study) will vary based on heterogeneity of and distribution of water masses. To inherently include satellite product uncertainties into the model, the network could alternatively be trained with satellite data as the input and imagery-based diatom carbon as the target. However, this would require an approach that defines the in situ-measured diatom carbon model target data on the spatial resolution of the satellite, in which assumptions must be made regarding spatial homogeneity of diatom populations.

5.5 Conclusions

Estimates of PCC on broad spatial scales in quantitative units and with associated uncertainties are needed for studies of biogeochemical cycles and ocean ecosystems and food webs. Pigments have been routinely used in numerous studies to predict distributions of diatoms and results from this study are the first to evaluate the assumptions made when using pigments and highlight the need for independent measurements of diatom carbon. The application of a trained neural network successfully models diatom carbon using input parameters that are currently measurable from satellites, making diatoms observable from space with significantly improved accuracy over what is possible using relationships between diatoms and Chl *a* alone. Propagation of uncertainties in model inputs and targets, as well as neural network errors, result in diatom carbon estimates with 65% median relative error. To our knowledge, all previously published algorithms to detect PCC from satellite do not account for the uncertainty in the in situ measurements used to develop the given algorithm. Algorithms developed without propagation of uncertainty in the measurements used for model development leave the user unable to determine the accuracy of a given method. The framework of training and applying neural networks is becoming more popular in earth science applications as more data become available to train machine-learning models. Further analysis is required to test for robustness and model applicability to times and regions differing from data used during model development. However, as methods such as quantitative cell imagery become more widely adopted, an increased number of datasets for model development and evaluation will be available. This advancement in data collected in situ, along with added information that will become available from hyperspectral ocean color satellite measurements in the near future (e.g., the upcoming NASA PACE mission) will enable estimates of PCC on broad ocean scales with greater detail and accuracy over what is currently possible. Knowledge on pigments alone is not sufficient to accurately predict diatom carbon and information on other environmental parameters should be considered.

CHAPTER 6

CONCLUSIONS AND FUTURE DIRECTIONS

Detecting phytoplankton community composition (PCC) on global and ocean basin scales has a wide range of applications in oceanographic research, both for addressing present-day questions as well as for accurate modeling of global biogeochemical cycles with variable future climate change scenarios. Looking forward, satellite-based information is critical for efforts to observe and understand changes in ocean ecosystems in the context of a changing climate (e.g., Rousseaux and Gregg 2015; Yang et al. 2020). Beyond bulk phytoplankton biomass, the importance of PCC is relevant in the context of a changing earth and ocean climate. Phytoplankton communities are changing at ocean basin scales due to shifts in environmental and physical conditions including increased ocean acidification (Hays et al. 2005; Petrou et al. 2019), regional decline in nutrients and decreased of mixed layer depth (Rousseaux and Gregg 2015), and northward expansion of warming temperate waters (Neukermans et al. 2018). Climate-driven increase in ocean stratification may lead to higher light and lower nutrients that in turn could increase the dominance of small phytoplankton on a global scale (Dutkiewicz et al. 2013). This could have a cascading effect on upper trophic level species, such as a shift towards more jellyfish and fewer pelagic fish in some parts of the world (Richardson et al. 2009; Uye 2011).

The remote nature of satellite measurements requires development of algorithms that link observations from space to PCC in ocean waters, as well as rigorous evaluation and validation of said algorithms. The research presented here was motivated at the outset by predicting PCC in the surface ocean from hyperspectral optical measurements. Chapter 3, and the preceding Master's thesis research, addressed this challenge by deriving concentrations of phytoplankton accessory pigments and pigment groups from hyperspectral optical measurements (Chase et al. 2013; 2017). Median errors reported for estimating phytoplankton accessory pigments both from spectral decomposition and using pigment covariation relationships were similar in value. The results of Chapter 3 not only provide an algorithm to

estimate pigments from hyperspectral $R_{rs}(\lambda)$, but also show that comparison of algorithms against what can be estimated directly from Chl *a* alone is necessary when using optical proxies to estimate PCC. One unanswered question following the work of Chapter 3 is how to best make use of optically-estimated pigments while accounting for underlying covariation between Chl *a* and accessory pigments (the calculation of a pigment anomaly as in Eq. 5.7 is one approach, for example).

Moving beyond the knowledge of pigments to define PCC is a non-trivial task. As evidenced by the data presented in Chapters 4 and 5 of this thesis, relationships between PCC and pigments are highly variable. Chapter 4 provides revised pigment-based phytoplankton size class equations, but perhaps more importantly, the work to evaluate the traditionally assigned pico-, nano-, and microplankton size classes using cytometry data brings to light the question of whether the three size classes are appropriate, as nearly all phytoplankton groups fall within the 2-20 μm nanoplankton class (Fig. 4.4; Table 4.4). The observation that a large diversity of phytoplankton types are found within all three size classes highlights the challenge in using diagnostic pigments to accurately define how a given phytoplankton community is distributed among the three classes. This prompts the question of whether the diagnostic pigment-based size classes warrant more drastic revision, for example in their size cutoffs, or perhaps in a revised approach to their interpretation in general.

Methods to quantify PCC such as quantitative cell imagery and flow cytometry, in combination with the high spatial and temporal resolution of optically-derived pigments, provide opportunities to measure pigments and PCC coincidently and observe the variability in pigment-PCC relationships over a wide range of oceanographic conditions (i.e., seasons and water types). The work of Chapter 5 suggests the need to define diatom carbon using methods other than pigments alone, as the relationships assumed between diagnostic pigments and diatoms do not capture the variability of diatoms as a function of pigments. Following the work of Chapter 5 (also see Appendix), a remaining questions is how to best

apply the knowledge gained from hyperspectral (vs. multispectral) satellite $R_{rs}(\lambda)$ measurements when relating optical and environmental parameters to diatom carbon concentrations. We found that the inclusion of optically-derived accessory pigments in a neural network approach to estimate diatom carbon improves the network results, but only marginally. This warrants further research and development of new approaches, as upcoming satellite missions will provide hyperspectral $R_{rs}(\lambda)$ at spatial and temporal resolutions not previously available.

6.1 Future Work

Further work to develop methods to predict PCC distributions from variables that can be estimated from space will be undertaken during my postdoctoral research will be focused in several areas:

1. Application of neural networks to predict PCC (e.g., as in Chapter 5), but with time-lag information that takes into account the history of water masses, as this is potentially an important driver of PCC distributions.
2. Incorporation of global data-assimilating models such as the Mercator Global Ocean forecasting system (GLO, Hernandez et al. 2015; Lellouche et al. 2018), which provides data including 3D temperature, salinity, vector currents, maps of sea surface height, and mixed layer depth. This type of information might be used in algorithms to predict mesoscale variability in PCC.
3. Further development of image classification networks for cell imagery at higher taxonomic resolution to provide a means of validation/comparison with remote sensing products.
4. Algorithms to predict PCC that make use of hyperspectral remote-sensing reflectance ($R_{rs}(\lambda)$) in support of upcoming hyperspectral ocean color satellite instruments, both by use of inversion for accessory pigment concentrations, and direct use of full spectra in neural network and machine learning applications.

Future work directions will also move beyond the development of algorithms to their application to answer outstanding questions about how phytoplankton communities are distributed, and how and why PCC changes in space and in time. For example, understanding phytoplankton species succession across seasons and over an annual cycle has been addressed using generalizations between broad phytoplankton groups and their nutrient requirements and environmental niches (Margalef 1978; Wyatt 2014). Recent studies consider the complexities of phytoplankton species succession and bloom formation (Glibert 2016; Kemp and Villareal 2018). Continued work related to this topic, such as a closer look at the succession and dynamics of different diatom species rather than considering diatoms as a uniform group, will improve our understating of complex relationships between PCC and the environment. Throughout all planned work, careful consideration of data uncertainties and limitations need to be included; see Chapter 2 for a discussion of how this pertains to hyperspectral measurements. One caveat to satellite-based observations is their limitation to surface ocean waters, where the depth of the "surface" is a function of the wavelength-dependent light attenuation properties of the water at a given time and place. Empirical relationships between surface measurements and vertical profiles of phytoplankton are one approach to addressing this challenge (e.g., Uitz et al. 2006).

6.2 Summary

The work conducted during this thesis began with deriving biologically relevant parameters from optical measurements, with the ultimate goal of detecting phytoplankton communities from space. Despite remaining challenges in predicting PCC from pigments derived from $a_p(\lambda)$ or $R_{rs}(\lambda)$, novel information was gained on relationships between pigments and phytoplankton size classes and diatom carbon. The fundamental knowledge of ocean optics gained during my graduate research and study will be the platform that algorithms for space-based PCC detection will build upon, and in a setting of interdisciplinary oceanographic research, the optics, biology, and physics approaches to understanding

PCC on broad ocean scales will be merged together. This thesis both presents novel methods, and adds knowledge to current understanding of the opportunities and challenges in linking pigments to PCC (Table 6.1). Future work will make use of the methods presented here as well as move beyond current approaches to explore new paths both for algorithm development, and application. All three research chapters (3, 4, and 5) contribute to the field of oceanography through the presentation of novel data and ideas, and the establishment of new questions and knowledge gaps to be addressed by future research.

Table 6.1. Summary of data types, approaches, and associated products. This table shows the results of analyses of the present doctoral thesis research, as well as in the preceding Master's thesis research.

Measurements		Approach	Products
Algorithm inputs	*Algorithm targets*		
$a_p(\lambda)$ spectra	HPLC pigments	Spectral decomposition	Accessory pigments *Chase et al. 2013*
$R_{rs}(\lambda)$ spectra	HPLC pigments	Spectral inversion & decomposition	Accessory pigments *Chase et al. 2017*
HPLC pigments	Combined cytometry data	Diagnostic Pigment Analysis	Phytoplankton Size Classes *Chase et al. in review*
Chl *a*, Temp, Salinity, day length, DOY	IFCB imagery	Neural Network	Diatom Carbon *Chase et al. in prep.*

REFERENCES

Aguirre-Gomez, R., A.R. Weeks, and S.R. Boxall. 2001. The Identification of Phytoplankton Pigments from Absorption Spectra. 22 (2): 315–38. https://doi.org/10.1080/014311601449952.

Alldredge, A., and M. Silver. 1988. Characteristics, Dynamics and Significance of Marine Snow. *Progress in Oceanography* 20: 41–82.

Alvain, S., C. Moulin, Y. Dandonneau, and F.M. Bréon. 2005. Remote Sensing of Phytoplankton Groups in Case 1 Waters from Global SeaWiFS Imagery. *Deep-Sea Res. Pt. I*, 52(2005), 1984 – 2004.

Alvain, S., C. Moulin, Y. Dandonneau, and H. Loisel. 2008. Seasonal Distribution and Succession of Dominant Phytoplankton Groups in the Global Ocean: A Satellite View, *Global Biogeochem. Cy.* 22(GB3001).

Alvain, S., H. Loisel, and D. Dessailly. 2012. Theoretical Analysis of Ocean Color Radiances Anomalies and Implications for Phytoplankton Groups Detection in Case 1 Waters. *Opt. Express* 20(2), 1070–83.

Antoine, D., A. Morel, H.R. Gordon, V.F. Banzon, and R.H. Evans. 2005. Bridging Ocean Color Observations of the 1980s and 2000s in Search of Long-Term Trends. *Journal of Geophysical Research* 110 (C06009). https://doi.org/10.1029/2004JC002620.

Armbrust, E.V. 2009. The Life of Diatoms in the World's Oceans. *Nature* 459 (7244): 185–92. https://doi.org/10.1038/nature08057.

Babin, M., and D. Stramski. 2004. Variations in the mass-specific absorption coefficient of mineral particles suspended in water, *Limnol. Oceanogr.* 49, 756–767.

Barton, A.D., M.S. Lozier, and R.G. Williams. 2015. Physical Controls of Variability in North Atlantic Phytoplankton Communities. *Limnol. Oceanog.* 60 (1): 181–97.

Bates, T.S., R.J. Charlson, and R.H. Gammon. 1987. Evidence for the Climatic Role of Marine Biogenic Sulphur. *Nature* 329 (6137): 319–21. https://doi.org/10.1038/329319a0.

Behrenfeld, M.J., J.T. Randerson, C.R. McClain, G.C. Feldman, S.O. Los, C.J. Tucker, P.G. Falkowski, et al. 2001. Biospheric Primary Production during an ENSO Transition. *Science* 291 (5513): 2594–97. https://doi.org/10.1126/science.1055071.

Behrenfeld, M.J., E. Boss, D.A. Siegel, and D.M. Shea. 2005. Carbon-Based Ocean Productivity and Phytoplankton Physiology from Space. *Global Biogeochemical Cycles* 19 (1). https://doi.org/10.1029/2004GB002299.

Behrenfeld, M.J., S.C. Doney, I. Lima, E.S. Boss, and D.A. Siegel. 2013. Annual Cycles of Ecological Disturbance and Recovery Underlying the Subarctic Atlantic Spring Plankton Bloom. *Global Biogeochemical Cycles* 27 (2): 526–40. https://doi.org/10.1002/gbc.20050.

Behrenfeld, M.J., R.T.O. Malley, E.S. Boss, T.K. Westberry, J.R. Graff, K.H. Halsey, A.J. Milligan, D.A. Siegel, and M.B. Brown. 2016. Revaluating Ocean Warming Impacts on Global Phytoplankton. *Nature Climate Change* 6: 323–30. doi:10.1038/NCLIMATE2838.

Behrenfeld, M. J., R. H. Moore, C. A. Hostetler and others. 2019. The North Atlantic Aerosol and Marine Ecosystem Study (NAAMES): Science Motive and Mission Overview, *Front. Mar. Sci.* 6:122, doi:10.3389/fmars.2019.00122.

Ben Mustapha, Z., S. Alvain, C. Jamet, H. Loisel, and D. Dessailly. 2013. Automatic Classification of Water-Leaving Radiance Anomalies from Global SeaWiFS Imagery: Application to the Detection of Phytoplankton Groups in Open Ocean Waters. *Remote Sens. Environ.* 146, 97-112.

Bidigare, R.R., M.E. Ondrusek, J.H. Morrow, and D.A. Kiefer. 1990. In-Vivo Absorption Properties of Algal Pigments. *Conference Proceedings, Ocean Optics X* SPIE 1302: 290–302.

Bopp, L. 2005. Response of Diatoms Distribution to Global Warming and Potential Implications: A Global Model Study. *Geophys. Res. Lett.* 32 (19): 2–5. doi:10.1029/2005GL023653.

Bopp, L., P. Monfray, O. Aumont, J.-L. Dufresne, H. L. Treut, G. Madec, L. Terray, and J. C. Orr. 2001. Potential Impact of Climate Change on Marine Export Production. *Global Biogeochem. Cy.* 15 (1): 81–99.

Boss, E., M. Picheral, T. Leeuw, A. Chase, E. Karsenti, G. Gorsky, L. Taylor, W. Slade, J. Ras, and H. Claustre. 2013. The Characteristics of Particulate Absorption, Scattering and Attenuation Coefficients in the Surface Ocean; Contribution of the Tara Oceans Expedition, *Methods in Oceanography*, 7, 52-62.

Boss, E., M. Picheral, W. Slade, L. Taylor, T. Leeuw, A. Chase. 2014. Tara Oceans Consortium, Coordinators; Tara Oceans Expedition, Participants: Properties of seawater and particulate matter from a WETLabs AC-S spectrophotometer and a WETLabs chlorophyll fluorometer mounted on the continuous surface water sampling system during the Tara Oceans expedition 2009-2013. doi:10.1594/PANGAEA.836318

Boyd, P., and P. Newton. 1995. Evidence of the Potential Influence of Planktonic Community Structure on the Interannual Variability of Particulate Organic Carbon Flux. *Deep-Sea Research Part I* 42 (5): 619–39. https://doi.org/10.1016/0967-0637(95)00017-Z.

Bracher, A., M. Vountas, T. Dinter, J.P. Burrows, R. Röttgers, and I. Peeken. 2009. Quantitative Observation of Cyanobacteria and Diatoms from Space Using PhytoDOAS on SCIAMACHY Data. *Biogeosciences* 6 (5): 751–64. https://doi.org/10.5194/bg-6-751-2009.

Bracher, A., M.H. Taylor, B. Taylor, T. Dinter, R. Röttgers, and F. Steinmetz. 2015. Using Empirical Orthogonal Functions Derived from Remote-Sensing Reflectance for the Prediction of Phytoplankton Pigment Concentrations, *Ocean Sci.,* 11(1), 139–58, doi:10.5194/os-11-139-2015.

Bracher, A., H. Bouman, R.J. Brewin, A. Bricaud, A.M. Ciotti, L. Clementson, E. Devred, et al. 2017. Obtaining Phytoplankton Diversity from Ocean Color : A Scientific Roadmap for Future Development, *Front. Mar. Sci.*, 4 (March): 1–15. doi:10.3389/fmars.2017.00055.

Brett, M.T, D.C. Muller-Navarra, and S. Park. 2000. Empirical Analysis of the Effect of Phosphorus Limitation on Algal Food Quality for Freshwater Zooplankton. *Limnol. Oceanogr.* 45 (7): 1564–75.

Brewin, R.J., N.J. Hardman-Mountford, S.J. Lavender and others. 2011. An intercomparison of bio-optical techniques for detecting dominant phytoplankton size class from satellite remote sensing, *Rem. Sens. Enviro.* 115(2): 325–339, doi:10.1016/j.rse.2010.09.004.

Brewin, R.J., S. Sathyendranath, T. Jackson, R. Barlow, V. Brotas, R. Airs, and T. Lamont. 2015. Influence of light in the mixed-layer on the parameters of a three-component model of phytoplankton size class. *Rem. Sens. Enviro.* 168: 437–450, doi:10.1016/j.rse.2015.07.004.

Brewin, R.J.W., S. Sathyendranath, T. Hirata, S.J. Lavender, R.M. Barciela, and N.J. Hardman-Mountford. 2010. A three-component model of phytoplankton size class for the Atlantic Ocean, *Ecol. Modell.* 221(11): 1472–1483, doi:10.1016/j.ecolmodel.2010.02.014.

Brewin, R.J.W., S. Sathyendranath, G. Tilstone, P.K. Lange, and T. Platt. 2014. A multicomponent model of phytoplankton size structure, *J. Geophys. Res.: Oceans.* 119: 1–19, doi:10.1002/2014JC009859.

Brewin, R.J. W., S. Ciavatta, S. Sathyendranath and others. 2017. Uncertainty in ocean-color estimates of chlorophyll for phytoplankton groups, *Front. Mar. Sci.* 4:104, doi:10.3389/fmars.2017.00104.

Brewin, R.J.W., G.H. Tilstone, T. Jackson, T. Cain, P.I. Miller, P.K. Lange, A. Misra, and R.L. Airs. 2016. Modelling Size-Fractionated Primary Production in the Atlantic Ocean from Remote Sensing. *Progress in Oceanography*. https://doi.org/10.1016/j.pocean.2017.02.002.

Bricaud, A., M. Babin, A. Morel, and H. Claustre. 1995. Variability in the Chlorophyll-Specific Absorption Coefficients of Natural Phytoplankton: Analysis and Parameterization, *J. Geophys. Res.,* 100(C7), 13,321–13,332.

Bricaud, A., A. Morel, M. Babin, K. Allali, and H. Claustre. 1998. Variations of Light Absorption by Suspended Particles with Chlorophyll a Concentration in Oceanic (Case 1) Waters: Analysis and Implications for Bio-Optical Models, *J. Geophys. Res.,* 103(C13), 31,033–31,044.

Bricaud, A., H. Claustre, J. Ras, K. Oubelkheir. 2004. Natural variability of phytoplanktonic absorption in oceanic waters: influence of the size structure of algal populations, *J. Geophys. Res.,* 109, C11010.

Bricaud, A., C. Mejia, D. Blondeau-Patissier, H. Claustre, M. Crepon, and S. Thiria. 2007. Retrieval of Pigment Concentrations and Size Structure of Algal Populations from Their Absorption Spectra Using Multilayered Perceptrons. *Applied Optics* 46 (8): 1251. doi:10.1364/AO.46.001251.

Brotas, V., R. J. Brewin, C. Sá and others. 2013. Deriving phytoplankton size classes from satellite data: Validation along a trophic gradient in the eastern Atlantic Ocean, *Rem. Sens. Enviro*. 134: 66–77, doi:10.1016/j.rse.2013.02.013.

Brown, C. A., Y. Huot, P. J. Werdell, B. Gentili, and H. Claustre. 2008. The Origin and Global Distribution of Second Order Variability in Satellite Ocean Color and Its Potential Applications to Algorithm Development, *Remote Sens. Environ.* 112(12), 4186–4203, doi:10.1016/j.rse.2008.06.008.

Campbell, J.W., and T. Aarup. 1992. New Production in the North Atlantic Derived from Seasonal Patterns of Surface Chlorophyll. *Deep-Sea Research* 39 (10): 1669–94.

Catlett, D., and D.A. Siegel. 2018. Phytoplankton Pigment Communities Can Be Modeled Using Unique Relationships With Spectral Absorption Signatures in a Dynamic Coastal Environment. *Journal of Geophysical Research: Oceans*, 246–64. https://doi.org/10.1002/2017JC013195.

Chase, A., E. Boss, R. Zaneveld, A. Bricaud, H. Claustre, J. Ras, G. Dall'Olmo, and T.K. Westberry. 2013. Decomposition of in Situ Particulate Absorption Spectra, *Methods in Oceanography*, 7, 110-124.

Chase, A.P., E. Boss, I. Cetinić, and W. Slade. 2017. Estimation of Phytoplankton Accessory Pigments From Hyperspectral Reflectance Spectra: Toward a Global Algorithm. *Journal of Geophysical Research: Oceans* 122 (12): 9725–43. doi:10.1002/2017JC012859.

Chase, A.P., S. Kramer, N. Haëntjens, E. Boss, L. Karp-Boss, M. Edmondson, J. Graff. 2020. Evaluation of diagnostic pigments to estimate phytoplankton size classes. *Limnol. Oceanog.: Methods*. *In revision*

Chisholm, S. W. 1992. Phytoplankton Size, p. 213–237. In P.G. Falkowski, A. D. Woodhead, and K. Vivirito [eds.], Primary Productivity and Biogeochemical Cycles in the Sea. Springer.

Ciotti, A.M., M.R. Lewis, and J.J. Cullen. 2002. Assessment of the Relationships between Dominant Cell Size in Natural Phytoplankton Communities and the Spectral Shape of the Absorption Coefficient. *Limnology and Oceanography* 47 (2): 404–17. https://doi.org/10.4319/lo.2002.47.2.0404.

Claustre, H. 1994. The trophic status of various oceanic provinces as revealed by phytoplankton pigment signatures. *Limnol. Oceanogr.* 39(5): 1206–1210.

Clementson, L.A., and B. Wojtasiewicz. 2019. Dataset on the Absorption Characteristics of Extracted Phytoplankton Pigments. *Data in Brief* 24: 103875. https://doi.org/10.1016/j.dib.2019.103875.

Clopper, C.J., and E.S. Pearson. 1934. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika* 26 (4): 404. https://doi.org/10.2307/2331986.

Craig, S.E., S.E. Lohrenz, Z. Lee, K.L. Mahoney, G.J. Kirkpatrick, O.M. Schofield, and R.G. Steward. 2006. Use of Hyperspectral Remote Sensing Reflectance for Detection and Assessment of the Harmful Alga, Karenia Brevis. *Applied Optics* 45: 5414–25. https://doi.org/10.1364/AO.45.005414.

Della Penna, A., and P. Gaube. 2019. Overview of (sub)mesoscale ocean dynamics for the NAAMES field program, *Front. Mar. Sci.* 6: 384, doi: 10.3389/fmars.2019.00384.

Delwiche, C.F. 1999. Tracing the Thread of Plastid Diversity through the Tapestry of Life. *The American Naturalist* 154 (S4): S164–77. doi:10.1086/303291.

Deng, L., W. Zhou, W. Cao, W. Zheng, G. Wang, Z. Xu, C. Li, Y. Yang, S. Hu, and W. Zhao. 2019. Retrieving Phytoplankton Size Class from the Absorption Coefficient and Chlorophyll A Concentration Based on Support Vector Machine. *Remote Sensing* 11 (9): 1054. https://doi.org/10.3390/rs11091054.

Devred, E., S. Sathyendranath, V. Stuart, and T. Platt. 2011. A three component classification of phytoplankton absorption spectra: Application to ocean-color data, *Rem. Sens. Enviro.* 115(9): 2255–2266, doi:10.1016/j.rse.2011.04.025.

Di Cicco, A., M. Sammartino, S. Marullo, and R. Santoleri. 2017. Regional empirical algorithms for an improved identification of phytoplankton functional types and size classes in the Mediterranean Sea using satellite data, *Front. Mar. Sci.* 4: 126, doi:10.3389/fmars.2017.00126.

Dickman, E.M., J.M. Newell, M.J. Gonza, and M.J. Vanni. 2008. Light, Nutrients, and Food-Chain Length Constrain Planktonic Energy Transfer Efficiency across Multiple Trophic Levels. *Proc. Nat. Acad. Sci.* 105 (47): 18408–12.

Doney, S.C., D.M. Glover, S.J. Mccue, and M. Fuentes. 2003. Mesoscale Variability of Sea-Viewing Wide Field-of-View Sensor (SeaWiFS) Satellite Ocean Color: Global Patterns and Spatial Scales. *J. Geophys. Res.* 108 (C2, 3024): 1–15. https://doi.org/10.1029/2001JC000843.

Dutkiewicz, S., M. Follows, J. Marshall, and W.W. Gregg. 2001. Interannual Variability of Phytoplankton Abundances in the North Atlantic. *Deep Sea Research Part II: Topical Studies in Oceanography* 48 (10): 2323–44. https://doi.org/10.1016/S0967-0645(00)00178-8.

Dutkiewicz, S., M.J. Follows, and J.G. Bragg. 2009. Modeling the Coupling of Ocean Ecology and Biogeochemistry. *Global Biogeochem. Cy.* 23 (4): 1–15. doi:10.1029/2008GB003405.

Dutkiewicz, S., J.R. Scott, and M.J. Follows. 2013. Winners and Losers: Ecological and Biogeochemical Changes in a Warming Ocean. *Global Biogeochem. Cy.* 27 (2): 463–77.

El Hourany, R., M. .A-Abi Saab, G. Faour, O. Aumont, M. Crépon, and S. Thiria. 2019. Estimation of Secondary Phytoplankton Pigments from Satellite Observations Using Self-Organizing Maps (SOM). *Journal of Geophysical Research: Oceans*. https://doi.org/10.1029/2018JC014450.

Ellen, J.S., C.A. Graff, and M.D. Ohman. 2019. Improving Plankton Image Classification Using Context Metadata. *Limnol. Oceanog.: Methods* 10324: lom3.10324.

Estapa, M.L., E. Boss, L.M. Mayer, and C.S. Roesler. 2012. Role of Iron and Organic Carbon in Mass-Specific Light Absorption by Particulate Matter from Louisiana Coastal Waters, *Limnol. Oceanogr.* 57 (1): 97–112. doi:10.4319/lo.2012.57.1.0097.

Eppley, R.W., and B.J. Peterson. 1979. Particulate Organic Matter Flux and Planktonic New Production in the Deep Ocean, *Nature* 282 (5740): 677–80. doi:10.1038/282677a0.

Farikou, O., S. Sawadogo, A. Niang, and D. Diouf. 2015. Inferring the Seasonal Evolution of Phytoplankton Groups in the Senegalo-Mauritanian Upwelling Region from Satellite Ocean-Color Spectral Measurements, *J. Geophys. Res.: Oceans* 1: 1–46. doi:10.1002/2014JC010023.

Ferreira, A., D. Stramski, C.A. E. Garcia, V.M. T. Garcia, Á.M. Ciotti, and C.R.B. Mendes. 2013. Variability in Light Absorption and Scattering of Phytoplankton in Patagonian Waters: Role of Community Size Structure and Pigment Composition. *Journal of Geophysical Research: Oceans* 118 (2): 698–714. https://doi.org/10.1002/jgrc.20082.

Field, C. B., M.J. Behrenfeld, J.T. Randerson, and P. Falkowski. 1998. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science* 281 (5374): 237–40. doi:10.1126/science.281.5374.237.

Follows, M.J., S. Dutkiewicz, S. Grant, and S.W. Chisholm. 2007. Emergent Biogeography of Microbial Communities in a Model Ocean. *Science*, http://www.sciencemag.org/content/315/5820/1843.short.

Fox, J., M.J. Behrenfeld, N. Haëntjens, A. Chase, S. J. Kramer, E. Boss, L. Karp-boss, et al. 2020. Phytoplankton Growth and Productivity in the Western North Atlantic: Observations of Regional Variability From the NAAMES Field Campaigns. *Frontiers in Marine Science* 7 (24): 1–15. https://doi.org/10.3389/fmars.2020.00024.

Gaube, P., C. Barceló, D.J. McGillicuddy, A. Domingo, P. Miller, B. Giffoni, N. Marcovaldi, and Y. Swimmer. 2017. The Use of Mesoscale Eddies by Juvenile Loggerhead Sea Turtles (Caretta Caretta) in the Southwestern Atlantic. *PLoS ONE* 12 (3). https://doi.org/10.1371/journal.pone.0172839.

Gillis, D.B., J.H. Bowles, M.J. Montes, and W.J. Moses. 2018. Propagation of Sensor Noise in Oceanic Hyperspectral Remote Sensing. *Optics Express* 26 (18): A818. https://doi.org/10.1364/OE.26.00A818.

Gitelson, A.A., B. Gao, R.-R. Li, S. Berdnikov, and V. Saprygin. 2011. Estimation of Chlorophyll-a Concentration in Productive Turbid Waters Using a Hyperspectral Imager for the Coastal Ocean—the Azov Sea Case Study. *Environmental Research Letters* 6 (2): 024023. https://doi.org/10.1088/1748-9326/6/2/024023.

Gittings, J. A., R. J. W. Brewin, D. E. Raitsos, M. Kheireddine, M. Ouhssain, B. H. Jones, and I. Hoteit. 2019. Remotely sensing phytoplankton size structure in the Red Sea, Rem. Sens. Enviro. 234: 111387, doi.org/10.1016/j.rse.2019.111387.

Glibert, P.M. 2016. Margalef Revisited: A New Phytoplankton Mandala Incorporating Twelve Dimensions, Including Nutritional Physiology. *Harmful Algae* 55: 25–30.

Goericke, R., and J.P. Montoya. 1998. Estimating the Contribution of Microalgal Taxa to Chlorophyll a in the Field - Variations of Pigment Ratios under Nutrient- and Light-Limited Growth. *Mar. Ecol. Prog. Ser.* 169: 97–112. https://doi.org/10.3354/meps169097.

Gordon, H. R., D.K. Clark, J.W. Brown, O.B. Brown, R.H. Evans, and W.W. Broenkow. 1983. Phytoplankton pigment concentrations in the Middle Atlantic Bight: Comparison of ship determinations and CZCS estimates, *Appl. Optics* 22, 3929-3931.

Gordon, H.R, and O.B. Brown. 1988. A Semianalytic Radiance Model of Ocean Color, *J. Geophys. Res.: Atmos.* 93 (D9), 10,909-10,924.

Gould, Jr., R.W., and G.A. Fryxell. 1988. Phytoplankton Species Composition and Abundance in a Gulf Stream Warm Core Ring. II. Distributional Patterns. *Journal of Marine Research* 46 (2): 399–428. https://doi.org/10.1357/002224088785113612.

Gregg, W.W., and K.L. Carder. 1990. A Simple Spectral Solar Irradiance Model for Cloudless Maritime Atmospheres, *Limnol. Oceanogr.* 35 (8), 1657–75.

Graff, J.R., and M.J. Behrenfeld. 2018. Photoacclimation responses in subarctic Atlantic phytoplankton following a natural mixing-restratification event. *Front. Mar. Sci.* 5: 209, doi:10.3389/fmars.2018.00209.

Guidi, L., L. Stemmann, G.A. Jackson, F. Ibanez, H. Claustre, L. Legendre, M. Picheral, and G. Gorsky. 2009. Effects of Phytoplankton Community on Production, Size, and Export of Large Aggregates: A World-Ocean Analysis. *Limnology and Oceanography* 54 (6): 1951–63. https://doi.org/10.4319/lo.2009.54.6.1951.

Guidi, L., S. Chaffron, L. Bittner, D. Eveillard, A. Larhlimi, S. Roux, Y. Darzi, et al. 2015. Plankton Networks Driving Carbon Export in the Oligotrophic Ocean, *Nature* 532, 465-470.

Guillou, L., M.J. Chrétiennot-Dinet, L.K. Medlin, H. Claustre, S. Loiseaux-de Goër, and D. Vaulot. 1999. Bolidomonas: A new genus with two species belonging to a new algal class, the Bolidophyceae Heterokonta*. J. Phycol.* 35(2): 368–381, doi: 10.1046/j.1529-8817.1999.3520368.x.

Hays, G.C., A.J. Richardson, and C. Robinson. 2005. Climate Change and Marine Plankton. *Trends in Ecology and Evolution* 20 (6): 337–44. https://doi.org/10.1016/j.tree.2005.03.004.

Hernandez, F., E. Blockley, G.B. Brassington, F. Davidson, P. Divakaran, M. Drévillon, S. Ishizaki, et al. 2015. Recent Progress in Performance Evaluations and near Real-Time Assessment of Operational Ocean Products. *Journal of Operational Oceanography* 8: s221–38. https://doi.org/10.1080/1755876X.2015.1050282.

Hirata, T., J. Aiken, N. Hardman-Mountford, T. Smyth, and R. Barlow. 2008. An absorption model to determine phytoplankton size classes from satellite ocean colour. *Rem. Sens. Enviro.* 112(6): 3153–3159, doi:10.1016/j.rse.2008.03.011.

Hirata, T., N.J. Hardman-Mountford, R.J.W. Brewin and others. 2011. Synoptic relationships between surface Chlorophyll-*a* and diagnostic pigments specific to phytoplankton functional types. *Biogeosciences.* 8(2): 311–327, doi:10.5194/bg-8-311-2011.

Hoepffner, N., and S. Sathyendranath. 1991. Effect of Pigment Composition on Absorption Properties of Phytoplankton, *Mar. Ecol. Prog. Ser.,* 73, 11–23.

Hoepffner, N., and S. Sathyendranath. 1993. Determination of the Major Groups of Phytoplankton Pigments from the Absorption Spectra of Total Particulate Matter, *J. Geophys. Res.: Oceans*, 98 (C12), 22,789–22,803.

Hoge, F.E., and P.E. Lyon. 1996. Satellite Retrieval of Inherent Optical Properties by Linear Matrix Inversion of Oceanic Models: An Analysis of Model and Radiance Measurement Errors. *Journal of Geophysical Research: Oceans* http://onlinelibrary.wiley.com/doi/10.1029/96JC01414/full.

Honjo, S., and S.J. Manganini. 1993. Annual Biogenic Particle Fluxes to the Interior of the North Atlantic Ocean; Studied at 34°N 21°W and 48°N 21°W. *Deep-Sea Res. Part II* 40 (1–2): 587–607. https://doi.org/10.1016/0967-0645(93)90034-K.

Hooker, S.B., L. Van Heukelem, C. S. Thomas and others. 2009. The third SeaWiFS HPLC analysis round-robin experiment SeaHARRE-3. NASA/TM-2009-215849. December, 2009.

Hu, C., Z. Lee, and B. Franz. 2012. Chlorophyll a Algorithms for Oligotrophic Oceans: A Novel Approach Based on Three-Band Reflectance Difference*, J. Geophys. Res.,* 117 (C1), C01011, doi:10.1029/2011JC007395.

Hunter-Cevera, K.R., M.G. Neubert, R.J. Olson, A.R. Solow, A. Shalapyonok, and H.M. Sosik. 2016. Physiological and Ecological Drivers of Early Spring Blooms of a Coastal Phytoplankter. *Science* 354 (6310): 326–29. https://doi.org/10.1126/science.aaf8536.

Huot, Y. and D. Antoine. 2016. Remote Sensing Reflectance Anomalies in the Ocean, *Remote Sens. Environ.* 184, 101–11, doi:10.1016/j.rse.2016.06.002.

IOCCG. 2014. Phytoplankton Functional Types from Space. Sathyendranath, S. (ed.), Reports ofthe International Ocean-Colour Coordinating Group, No. 15, IOCCG, Dartmouth, Canada.

Irigoien, X., R.P. Harris, H.M. Verheye, P. Joly, J. Runge, M. Starr, D. Pond, et al. 2002. Copepod Hatching Success in Marine Ecosystems with High Diatom Concentrations. *Nature* 419 (6905): 387–89. https://doi.org/10.1038/nature01055.

Isada, T., T. Hirawake, T. Kobayashi, Y. Nosaka, and M. Natsuike. 2015. Remote Sensing of Environment Hyperspectral Optical Discrimination of Phytoplankton Community Structure in Funka Bay and Its Implications for Ocean Color Remote Sensing of Diatoms, *Remote Sens. Environ.* 159, 134-151.

Iturriaga, R. and D.A. Siegel. 1989. Microphotometric characterization of phytoplankton and detrital absorption properties in the Sargasso Sea, *Limnol. Oceanogr.* 34, 1706–1726.

Jeffrey, V.M., R.F.C. Mantoura, and S.W. Morgan. 1997. Phytoplankton pigments in oceanography. UNESCO.

Jeffrey, S.W. and M. Vesk. 1997. Introduction to marine phytoplankton and their pigment signatures, in: Phytoplankton pigment in oceanography: Guidelines to modern methods, edited by: Jeffrey, S.W., R.F. C. Mantoura, and S.W. Wright, UNESCO, 33–84.

Kemp, A.E.S., and T.A. Villareal. 2018. The Case of the Diatoms and the Muddled Mandalas: Time to Recognize Diatom Adaptations to Stratified Waters. *Progress in Oceanography* 167 (August): 138–49. https://doi.org/10.1016/j.pocean.2018.08.002.

Kirk, J.T.O., 1994. Light and Photosynthesis in Aquatic Ecosystems, Second Edition, Cambridge Univ. Press.

Kirkpatrick, G., D.F. Millie, M.A. Moline, O. Schofield. 2000. Optical Discrimination of a Phytoplankton Species in Natural Mixed Populations. *Limnology and Oceanography* 45 (2): 467–71. http://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1139&context=bio_fac.

Kramer, S.J., C.S. Roesler, and H.M. Sosik. 2018. Bio-Optical Discrimination of Diatoms from Other Phytoplankton in the Surface Ocean: Evaluation and Refinement of a Model for the Northwest Atlantic. *Remote Sensing of Environment* 217 (August): 126–43. https://doi.org/10.1016/j.rse.2018.08.010.

Kramer, S.J., and D.A. Siegel. 2019. How Can Phytoplankton Pigments Be Best Used to Characterize Surface Ocean Phytoplankton Groups for Ocean Color Remote Sensing Algorithms? *Journal of Geophysical Research: Oceans*, 0–3. https://doi.org/10.1029/2019JC015604.

Kostadinov, T.S., D.A. Siegel, and S. Maritorena. 2009. Retrieval of the particle size distribution from satellite ocean color observations*. J. Geophys. Res.: Oceans.* 114(9): 1–22, doi:10.1029/2009JC005303.

Kostadinov, T.S., D.A. Siegel, and S. Maritorena. 2010. Global variability of phytoplankton functional types from space: assessment via the particle size distribution. *Biogeosciences*. 7(10): 3239–3257, doi:10.5194/bg-7-3239-2010.

Kuwata, A., K. Yamada, M. Ichinomiya, S. Yoshikawa, M. Tragin, D. Vaulot, and A.L. dos Santos. 2018. Bolidophyceae, a sister picoplanktonic group of diatoms - A review*, Front. Mar. Sci.* 5: 370, doi:10.3389/fmars.2018.00370.

Laney, S.R., and H.M. Sosik. 2014. Phytoplankton Assemblage Structure in and around a Massive Under-Ice Bloom in the Chukchi Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* 105 (July): 30–41. https://doi.org/10.1016/j.dsr2.2014.03.012.

Leblanc, K., J. Arístegui, L. Armand, P. Assmy, B. Beker, A. Bode, E. Breton, et al. 2012. A Global Diatom Database- A Bundance, Biovolume and Biomass in the World Ocean. *Earth System Science Data* 4 (1): 149–65. https://doi.org/10.5194/essd-4-149-2012.

Le Quere, C., S.P. Harrison, I.C. Prentice, E.T. Buitenhuis, O. Aumont, L. Bopp, H. Claustre, et al. 2005. Ecosystem Dynamics Based on Plankton Functional Types for Global Ocean Biogeochemistry Models. *Global Change Biology* 11: 2016–40. https://doi.org/10.1111/j.1365-2486.2005.01004.

Leathers, R., T.V. Downes, and C. Mobley. 2001. Self-Shading Correction for Upwelling Sea-Surface Radiance Measurements Made with Buoyed Instruments, *Opt. Express* 8 (10), 561-570. doi:10.1364/OE.8.000561.

Leblanc, K., J. Arístegui, L. Armand, P. Assmy, B. Beker, A. Bode, E. Breton, et al. 2012. A Global Diatom Database- A Bundance, Biovolume and Biomass in the World Ocean. *Earth System Science Data* 4 (1): 149–65. https://doi.org/10.5194/essd-4-149-2012.

Leblanc, K., B. Quéguiner, F. Diaz and others. 2018. Nanoplanktonic diatoms are globally overlooked but play a role in spring blooms and carbon export. *Nat. Commun.* 9(953): 1–12, doi:10.1038/s41467-018-03376-9.

Lee, Z., K.L. Carder, and R.A. Arnone. 2002. Deriving Inherent Optical Properties from Water Color: A Multiband Quasi-Analytical Algorithm for Optically Deep Waters, *Applied Optics,* 41 (27): 5755. doi:10.1364/AO.41.005755.

Lee, Z., K. Carder, R. Arnone, and M. He. 2007. Determination of Primary Spectral Bands for Remote Sensing of Aquatic Environments, *Sensors,* 7 (12): 3428–41. doi:10.3390/s7123428.

Lee, Z., K. Du, K.J. Voss, G. Zibordi, B. Lubac, R. Arnone, and A. Weidemann. 2011. An Inherent-Optical-Property-Centered Approach to Correct the Angular Effects in Water-Leaving Radiance, *Applied Optics,* 50 (19): 3155–67.

Lellouche, J.M., E. Greiner, O. Le Galloudec, G. Garric, C. Regnier, M. Drevillon, M. Benkiran, et al. 2018. Recent Updates to the Copernicus Marine Service Global Ocean Monitoring and Forecasting Real-Time 1g 12° High-Resolution System. *Ocean Science* 14 (5): 1093–1126. https://doi.org/10.5194/os-14-1093-2018.

Lévy, M., P.J. S. Franks, and K. Shafer Smith. 2018. The Role of Submesoscale Currents in Structuring Marine Ecosystems. *Nature Communications* 9 (1): 4758. https://doi.org/10.1038/s41467-018-07059-3.

Lévy, M., O. Jahn, S. Dutkiewicz, and M.J. Follows. 2014. Phytoplankton Diversity and Community Structure Affected by Oceanic Dispersal and Mesoscale Turbulence. *Limnology and Oceanography: Fluids and Environments* 4 (1): 67–84. https://doi.org/10.1215/21573689-2768549.

Li, Z., L. Li, K. Song, and N. Cassar. 2013. Estimation of phytoplankton size fractions based on spectral features of remote sensing ocean color data. *J. Geophys. Res.: Oceans.* 118(3): 1445–1458, doi:10.1002/jgrc.20137.

Liu, Y., E. Boss, A. Chase, H. Xi, X. Zhang, R. Röttgers, Y. Pan, and A. Bracher. 2019. Retrieval of Phytoplankton Pigments from Underway Spectrophotometry in the Fram Strait. 1–32. https://doi.org/10.3390/rs11030318.

Litchman, E., and C.A. Klausmeier. 2008. Trait-based community ecology of phytoplankton. *Ann. Rev. Ecol. Evol. Syst.* 39(1): 615–639, doi: 10.1146/annurev.ecolsys.39.110707.173549.

Litchman, E., P. de Tezanos Pinto, C.A. Klausmeier, M.K. Thomas, and K. Yoshiyama. 2010. Linking traits to species diversity and community structure in phytoplankton. *Hydrobiologia*. 653: 15–28, doi:10.1007/s10750-010-0341-5.

Lombard, F., E. Boss, A.M. Waite and others. 2019. Globally consistent quantitative observations of planktonic ecosystems, *Front. Mar. Sci.* 6:196, doi:10.3389/fmars.2019.00196.

Lohrenz, S.E., A. Weidemann, and M. Tuel. 2003. Phytoplankton Spectral Absorption as Influenced by Community Size Structure and Pigment Composition, *J. Plankton Res.,* 25 (1), 35–61.

Lorenzoni, L., G. Toro-Farmer, R. Varela, L. Guzman, J. Rojas, E. Montes, and F. Muller-Karger. 2015. Characterization of phytoplankton variability in the Cariaco Basin using spectral absorption, taxonomic and pigment data. *Rem. Sens. Enviro.* 167: 259–268, doi:10.1016/j.rse.2015.05.002.

Losa, S.N., M.A. Soppa, T. Dinter, A. Wolanin, R.J.W. Brewin, A. Bricaud, J. Oelker, et al. 2017. Synergistic Exploitation of Hyper- and Multi-Spectral Precursor Sentinel Measurements to Determine Phytoplankton Functional Types (SynSenPFT). *Frontiers in Marine Science* 4 (JUL): 1–22.

Lubac, B., H. Loisel, N. Guiselin, R. Astoreca, L. F. Artigas, and X. Mériaux. 2008. Hyperspectral and Multispectral Ocean Color Inversions to Detect Phaeocystis Globosa Blooms in Coastal Waters. *Journal of Geophysical Research: Oceans* 113 (6): 1–17. https://doi.org/10.1029/2007JC004451.

Lucke, R.L., M. Corson, N.R. McGlothlin, S.D Butcher, D.L. Wood, D.R. Korwan, R.R. Li, W.A. Snyder, C.O. Davis, and D.T. Chen. 2011. Hyperspectral Imager for the Coastal Ocean: Instrument Description and First Images. *Applied Optics* 50 (11): 1501–16. http://www.ncbi.nlm.nih.gov/pubmed/21478922.

Mackey, M., D. Mackey, H. W. Higgins, and S.W. Wright. 1996. CHEMTAX - a program for estimating class abundances from chemical markers: application to HPLC measurements of phytoplankton. *Mar. Ecol. Progr. Ser.* 144: 265–283.

Mao, Z., V. Stuart, D. Pan, J. Chen, F. Gong, H. Huang, and Q. Zhu. 2010. Effects of Phytoplankton Species Composition on Absorption Spectra and Modeled Hyperspectral Reflectance. *Ecological Informatics* 5 (5): 359–66. https://doi.org/10.1016/j.ecoinf.2010.04.004.

Margalef, R.. 1978. Life-Forms of Phytoplankton as Survival Alternatives in an Unstable Environment. *Oceanologica Acta* 1: 493–509.

Mason, J. D., M. T. Cone, and E. S. Fry. 2016. Ultraviolet (250–550 Nm) Absorption Spectrum of Pure Water, *Appl. Optics,* 55 (25), 7163, doi:10.1364/AO.55.007163.

Mckinna, L.I.W., P.J. Werdell, and C.W. Proctor. 2016. Implementation of an Analytical Raman Scattering Correction for Satellite Ocean-Color Processing, *Opt. Express,* 24 (14), 4241–55, doi:10.1364/OE.24.0A1123.

McKinna, L. I. W., I. Cetinić, A.P. Chase, and P.J. Werdell. 2019. Approach for Propagating Radiometric Data Uncertainties Through NASA Ocean Color Algorithms. *Frontiers in Earth Science* 7 (July): 1–17. https://doi.org/10.3389/feart.2019.00176.

Menden-Deuer, S., and E. J. Lessard. 2000. Carbon to volume relationships for dinoflagellates, diatoms, and other protist plankton. *Limnol. Oceanogr.* 45(3): 569-579.

Michaels, A.F., and M.W. Silver. 1988. Primary Production, Sinking Fluxes and the Microbial Food Web. *Deep-Sea Research* 35 (4): 473–90.

Moberg, E.A., and H.M. Sosik. 2012. Distance maps to estimate cell volume from two- dimensional plankton images, *Limnol. Oceanogr.: Methods*. 10: 278–288, doi:10.4319/lom.2012.10.278.

Mobley, C.D. 1994. Light and Water: Radiative Transfer in Natural Waters, *Academic*, San Diego, CA, USA.

Moisan, J.R., T. A.H. Moisan, and M.A. Linkswiler. 2011. An Inverse Modeling Approach to Estimating Phytoplankton Pigment Concentrations from Phytoplankton Absorption Spectra, *J. Geophys. Res.,* 116 (C9), C09018, doi:10.1029/2010JC006786.

Moon-Van Der Staay, S.Y., G.W. Van Der Staay, L. Guillou, D. Vaulot, H. Claustre, and L.K. Médlin. 2000. Abundance and diversity of prymnesiophytes in the picoplankton community from the equatorial Pacific Ocean inferred from 18S rDNA sequences. *Limnol. Oceanogr.* 45(1): 98–109, doi:10.4319/lo.2000.45.1.0098.

Morel, A., and A. Bricaud. 1981. Theoretical Results Concerning Light Absorption in a Discrete Medium, and Application to Specific Absorption of Phytoplankton. *Deep Sea Research Part A. Oceanographic Research* 28A (11): 1375–93. http://www.sciencedirect.com/science/article/pii/019801498190039X.

Morel, A., and B. Gentili. 1996. Diffuse Reflectance of Oceanic Waters. III. Implication of Bidirectionality for the Remote-Sensing Problem, *Appl. Optics* 35 (24), 4850–62, doi:10.1364/AO.35.004850.

Morel, A., Y. Huot, B. Gentili, P.J. Werdell, S.B. Hooker, and B.A. Franz. 2007. Examining the Consistency of Products Derived from Various Ocean Color Sensors in Open Ocean (Case 1) Waters in the Perspective of a Multi-Sensor Approach, *Remote Sens. Environ.* 111 (1), 69–88, doi:10.1016/j.rse.2007.03.012.

Mouw, C.B., A. Barnett, G.A. Mckinley, L. Gloege, and D. Pilcher. 2016. Phytoplankton Size Impact on Export Flux in the Global Ocean*, Global Biogeochem. Cy.* 30: 1542–62. doi:10.1002/2015GB005355.

Mouw, C.B., N.J. Hardman-Mountford, S. Alvain and others. 2017. A consumer's guide to satellite remote sensing of multiple phytoplankton groups in the global ocean. *Front. Mar. Sci.* 4: 41, doi:10.3389/fmars.2017.00041.

Nair, A., S. Sathyendranath, T. Platt, J. Morales, V. Stuart, M.-H. Forget, E. Devred, and H. Bouman. 2008. Remote Sensing of Phytoplankton Functional Types. *Remote Sensing of Environment* 112 (8): 3366–75. https://doi.org/10.1016/j.rse.2008.01.021.

Nelson, D.M., P.Tréguer, M.A. Brzezinski, A. Leynaert, and B. Quéguiner. 1995. Production and Dissolution of Biogenic Silica in the Ocean: Revised Global Estimates, Comparison with Regional Data and Relationship to Biogenic Sedimentation. *Global Biogeochem. Cy.* 9 (3): 359. https://doi.org/10.1029/95GB01070.

Neukermans, G., L. Oziel, and M. Babin. 2018. Increased Intrusion of Warming Atlantic Water Leads to Rapid Expansion of Temperate Phytoplankton in the Arctic. *Global Change Biology* 00: 1–9. https://doi.org/10.1111/gcb.14075.

Olson, R.J., and H.M. Sosik. 2007. A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot. *Limnol. Oceanogr.: Methods.* 5: 195–203, doi:10.4319/lom.2007.5.195.

O'Reilly, J.E., and 24 Coauthors. 2000. SeaWiFS Postlaunch Calibration and Validation Analyses, Part 3. NASA Tech. Memo. 2000-206892, Vol. 11, S.B. Hooker and E.R. Firestone, Eds., NASA Goddard Space Flight Center, 49 pp.

Organelli, E., A. Bricaud, D. Antoine, and J. Uitz. 2013. Multivariate approach for the retrieval of phytoplankton size structure from measured light absorption spectra in the Mediterranean Sea (BOUSSOLE site). *Appl. Opt.* 52(11): 2257–73.

Organelli, E., C. Nuccio, L. Lazzara, J. Uitz, A. Bricaud, and L. Massi. 2017. On the Discrimination of Multiple Phytoplankton Groups from Light Absorption Spectra of Assemblages with Mixed Taxonomic Composition and Variable Light Conditions. *Applied Optics* 56 (14): 3952. https://doi.org/10.1364/AO.56.003952.

Palacz, A.P., M.A. St. John, R.J.W. Brewin, T. Hirata, and W.W. Gregg. 2013. Distribution of Phytoplankton Functional Types in High-Nitrate, Low-Chlorophyll Waters in a New Diagnostic Ecological Indicator Model. *Biogeosciences* 10 (11): 7553–74. https://doi.org/10.5194/bg-10-7553-2013.

Pan, X., A. Mannino, M.E. Russ, S.B. Hooker, and L.W. Harding. 2010. Remote Sensing of Phytoplankton Pigment Distribution in the United States Northeast Coast, *Remote Sens. Environ.* 114 (11), doi:10.1016/j.rse.2010.05.015.

Petrou, K., K.G. Baker, D.A. Nielsen, A.M. Hancock, K.G. Schulz, and A.T. Davidson. 2019. Acidification Diminishes Diatom Silica Production in the Southern Ocean. *Nature Climate Change* 9 (10): 781–86.

Picheral, M., S. Searson, V. Taillandier, A. Bricaud, E. Boss, J. Ras, H. Claustre, M. Ouhssain, P. Morin, J-E. Tremblay, L. Coppola, J-P. Gattuso, N. Metzl, D. Thuillier, G. Gorsky, Gabriel. 2014. Tara Oceans Consortium, Coordinators; Tara Oceans Expedition, Participants: Vertical profiles of environmental parameters measured on discrete water samples collected with Niskin bottles during the Tara Oceans expedition 2009-2013. doi:10.1594/PANGAEA.836319.

Pope, R.M., and E.S. Fry. 1997. Absorption Spectrum (380–700 Nm) of Pure Water. II. Integrating Cavity Measurements, *Appl. Optics* 36 (33).

Poulin, C., D. Antoine, and Y. Huot. 2018. Diurnal Variations of the Optical Properties of Phytoplankton in a Laboratory Experiment and Their Implication for Using Inherent Optical Properties to Measure Biomass. *Optics Express* 26 (2): 711. https://doi.org/10.1364/OE.26.000711.

Raitsos, D.E., S.J. Lavender, C.D. Maravelias, J. Haralabous, A. J. Richardson, and P. C. Reid. 2008. Identifying Four Phytoplankton Functional Types from Space: An Ecological Approach, *Limnol. Oceanogr.* 53 (2): 605–13.

Ras, J., H. Claustre, and J. Uitz. 2008. Spatial Variability of Phytoplankton Pigment Distributions in the Subtropical South Pacific Ocean: Comparison between in Situ and Predicted Data, *Biogeosciences* 5, 353–69.

Richardson, A.J., A. Bakun, G.C. Hays, and M. J. Gibbons. 2009. The Jellyfish Joyride: Causes, Consequences and Management Responses to a More Gelatinous Future. *Trends in Ecology and Evolution* 24 (6): 312–22. https://doi.org/10.1016/j.tree.2009.01.010.

Roelke, D.L., C.D. Kennedy, and A.D. Weidemann. 1999. Use of Discriminant and Fourth-Derivative Analyses with High-Resolution Absorption Spectra for Phytoplankton Research: Limitations at Varied Signal-to-Noise Ratio and Spectral Resolution. *Gulf of Mexico Science* 17 (2): 75–86. https://doi.org/10.18785/goms.1702.02.

Roesler, C.S., M.J. Perry, and K.L. Carder. 1989. Modeling in Situ Phytoplankton Absorption from Total Absorption Spectra in Productive Inland Marine Waters, *Limnol.Oceanogr.* 34 (8), 1510–23, doi:10.4319/lo.1989.34.8.1510.

Roesler, C.S., and E. Boss. 2003. Spectral Beam Attenuation Coefficient Retrieved from Ocean Color Inversion*, Geophys. Res. Lett.* 30 (9), 1468, doi:10.1029/2002GL016185.

Roesler, C. S., S. M. Etheridge, and G.C. Pitcher. 2004. Application of an Ocean Color Algal Taxa Detection Model to Red Tides in the Southern Benguela. *Harmful Algae 2002*, 303–5.

Roesler, C.S., and M.J. Perry. 1995. In Situ Phytoplankton Absorption, Fluorescence Emission, and Particulate Backscattering Spectra Determined from Reflectance. *Journal of Geophysical Research:* 100 (C7): 13,279-13,294. http://onlinelibrary.wiley.com/doi/10.1029/95JC00455/full.

Rousseaux, C.S., and W.W. Gregg. 2015. Recent Decadal Trends in Global Phytoplankton Composition. *Global Biogeochemical Cycles* 29: 1674–88. https://doi.org/10.1002/2015GB005139.Received.

Roy, S., C.A. Llewellyn, E.S. Egeland, and G. Johnsen. 2011. Phytoplankton Pigments: Characterization, Chemotaxonomy and Applications in Oceanography. Cambridge University Press.

Roy, S., S. Sathyendranath, H. Bouman, and T. Platt. 2013. The global distribution of phytoplankton size spectrum and size classes from their light-absorption spectra derived from satellite data, *Rem. Sens. Enviro.* 139, 185–197, doi:10.1016/j.rse.2013.08.004.

Ryan, J.P., C. O. Davis, N.B. Tufillaro, R.M. Kudela, and B. C. Gao. 2014. Application of the Hyperspectral Imager for the Coastal Ocean to Phytoplankton Ecology Studies in Monterey Bay, CA, USA. *Remote Sensing* 6 (2): 1007–25. https://doi.org/10.3390/rs6021007.

Sabine, C.L., R.A. Feely, N. Gruber, R.M. Key, K. Lee, J.L. Bullister, R. Wanninkhof, et al. 2004. The Oceanic Sink for Anthropogenic CO2. *Science* 305 (5682): 367–71.

Sadeghi, A., T. Dinter, M. Vountas, B. Taylor, M. Altenburg-Soppa, and A. Bracher. 2012. Remote Sensing of Coccolithophore Blooms in Selected Oceanic Regions Using the PhytoDOAS Method Applied to Hyper-Spectral Satellite Data. *Biogeosciences* 9 (6): 2127–43.

Sanchez-Puerta, M.V., and C.F. Delwiche. 2008. A Hypothesis for Plastid Evolution in Chromalveolates. *Journal of Phycology* 44 (5): 1097–1107. https://doi.org/10.1111/j.1529-8817.2008.00559.x.

Sathyendranath, S., L. Watts, E. Devred, T. Platt, C. Caverhill, and H. Maass. 2004. Discrimination of Diatoms from Other Phytoplankton Using Ocean-Colour Data. *Marine Ecology Progress Series* 272: 59–68. https://doi.org/10.3354/meps272059.

Sathyendranath, S., V. Stuart, A. Nair, K. Oka, T. Nakane, H. Bouman, M.H. Forget, H. Maass, and T. Platt. 2009. Carbon-to-Chlorophyll Ratio and Growth Rate of Phytoplankton in the Sea. *Marine Ecology Progress Series* 383: 73–84. https://doi.org/10.3354/meps07998.

Seegers, B.N., R.P. Stumpf, B.A. Schaeffer, K.A. Loftin, and P. J. Werdell. 2018. Performance Metrics for the Assessment of Satellite Data Products: An Ocean Color Case Study. *Optics Express* 26 (6): 7404. https://doi.org/10.1364/oe.26.007404.

Shaju, S.S., P. Minu, A. S. Srikanth, P. Muhamed Ashraf, A. K. Vijayan, and B. Meenakumari. 2015. Decomposition Study of in Vivo Phytoplankton Absorption Spectra Aimed at Identifying the Pigments and the Phytoplankton Group in Complex Case 2 Coastal Waters of the Arabian Sea. *Oceanological and Hydrobiological Studies* 44 (3): 282–93. https://doi.org/10.1515/ohs-2015-0027.

Sheldon, R.W., A. Prakash, and H. Sutcliffe. 1972. The size distribution of properties in the ocean. *Limnol. Oceanogr.* 17(3): 327-340.

Siegel, D.A., K.O. Buesseler, S.C. Doney, S.F. Sailley, M.J. Behrenfeld, and P.W. Boyd. 2014. Global Assessment of Ocean Carbon Export by Combining Satellite Observations and Food-Web Models. *Global Biogeochemical Cycles* 301–14. https://doi.org/10.1002/2014GB004872.Received.

Slade, W.H., E. Boss, G. Dall'olmo, M.R. Langner, J. Loftin, M.J. Behrenfeld, C. Roesler, and T.K. Westberry. 2010. Underway and Moored Methods for Improving Accuracy in Measurement of Spectral Particulate Absorption and Attenuation, *J. Atmos. Oce. Technol.* 27 (10), 1733–46, doi:10.1175/2010JTECHO755.1.

Sieburth, J.M., V. Smetacek, and J. Lenz. 1978. Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions. *Limnol. Oceanogr.* 23(6): 1256–1263, doi:10.4319/lo.1978.23.6.1256.

Sommer, U., K. Lengfellner, and A. Lewandowska. 2012. Experimental induction of a coastal spring bloom early in the year by intermittent high-light episodes. *Mar. Ecol.: Prog. Ser.* 446: 61–71, doi:10.3354/meps09486.

Sommer, U., E. Charalampous, S. Genitsaris, and M. Moustaka-Gouni. 2017. Benefits, costs and taxonomic distribution of marine phytoplankton body size*. J. Plankton. Res.* 39(3): 494–508, doi:10.1093/plankt/fbw071.

Soppa, M.A., T. Hirata, B. Silva, T. Dinter, I. Peeken, S. Wiegmann, and A. Bracher. 2014. Global retrieval of diatom abundance based on phytoplankton pigments and satellite data. *Rem. Sens.* 6(10): 10,089–10,106, doi:10.3390/rs61010089.

Sosik, H.M., and B.Greg Mitchell. 1995. Light Absorption by Phytoplankton, Photosynthetic Pigments and Detritus in the California Current System. *Deep Sea Research Part I: Oceanographic Research Papers* 42 (10): 1717–48. https://doi.org/10.1016/0967-0637(95)00081-G.

Sosik, H.M., and R.J. Olson. 2007. Automated taxonomic classification of phytoplankton sampled with imaging in-flow cytometry. *Limnol. Oceanogr.: Methods.* 5: 204–216, doi:10.4319/lom.2007.5.204.

Suess, E. 1980. Particulate Organic Carbon Flux in the Oceans: Surface and Oxygen Utilization. *Nature* 288: 260–63.

Sullivan, J.M., M.S. Twardowski, J. R.V. Zaneveld, C.M. Moore, A.H. Barnard, P.L. Donaghay, and B. Rhoades. 2006. Hyperspectral Temperature and Salt Dependencies of Absorption by Water and Heavy Water in the 400-750 Nm Spectral Range, *Appl. Optics* 45 (21), 5294. doi:10.1364/AO.45.005294.

Swan, C.M., M. Vogt, N. Gruber, and C. Laufkoetter. 2016. A Global Seasonal Surface Ocean Climatology of Phytoplankton Types Based on CHEMTAX Analysis of HPLC Pigments. *Deep-Sea Research Part I: Oceanographic Research Papers* 109: 137–56. https://doi.org/10.1016/j.dsr.2015.12.002.

Tomas, C. R. 1997. Identifying Marine Phytoplankton. Academic Press.

Torrecilla, E., D. Stramski, R. Reynolds, Ed. Millán-Núñez, and J. Piera. 2011. Cluster Analysis of Hyperspectral Optical Data for Discriminating Phytoplankton Pigment Assemblages in the Open Ocean, *Remote Sens. Environ.* 115 (10), 2578–93. doi:10.1016/j.rse.2011.05.014.

Trees, C.C., D.K. Clark, R.R. Bidigare, M.E. Ondrusek, and J.L. Mueller. 2000. Accessory Pigments versus Chlorophyll a Concentrations within the Euphotic Zone: A Ubiquitous Relationship, *Limnol. Oceanogr.* 45 (5), 1130–43, doi:10.4319/lo.2000.45.5.1130.

Tréguer, P., D.M. Nelson, A.J. Van Bennekom, D.J. Demaster, A. Leynaert, and B. Quéguiner. 1995. The Silica Balance in the World Ocean: A Reestimate. *Science* 268 (5209): 375–79. https://doi.org/10.1126/science.268.5209.375.

Turner, J.T. 2015. Zooplankton Fecal Pellets, Marine Snow, Phytodetritus and the Ocean's Biological Pump. *Progress in Oceanography* 130: 205–48.

Twardowski, M.S., E. Boss, J.B. Macdonald, W. S. Pegau, A.H. Barnard, and J.R.V. Zaneveld. 2001. A Model for Estimating Bulk Refractive Index from the Optical Backscattering Ratio and the Implications for Understanding Particle Composition in Case I and Case II Waters*, J. Geophys. Res.* 106 (C7), 14129, doi:10.1029/2000JC000404.

Uitz, J., H. Claustre, A. Morel, and S.B. Hooker. 2006. Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *J. Geophys. Res.* 111: C08005, doi:10.1029/2005JC003207.

Uitz, J., Y. Huot, F. Bruyant, M. Babin, and H. Claustre. 2008. Relating phytoplankton photophysiological properties to community structure on large scales. *Limnol. Oceanogr.* 53(2): 614–630, doi:10.4319/lo.2008.53.2.0614.

Uitz, J., D. Stramski, R.A. Reynolds, and J. Dubranna. 2015. Assessing Phytoplankton Community Composition from Hyperspectral Measurements of Phytoplankton Absorption Coefficient and Remote-Sensing Reflectance in Open-Ocean Environments, *Remote Sens. Environ.* 171, 58–74.

van de Poll, W.H., G. Kulk, K.R. Timmermans, C.P.D. Brussaard, H.J. Van Der Woerd, M.J. Kehoe, K.D.A. Mojica, R.J.W. Visser, P.D. Rozema, and A.G.J. Buma. 2013. Phytoplankton Chlorophyll *a* Biomass, Composition, and Productivity along a Temperature and Stratification Gradient in the Northeast Atlantic Ocean. *Biogeosciences* 10 (6): 4227–40. https://doi.org/10.5194/bg-10-4227-2013.

Van Heukelem, L., and C.S. Thomas. 2001. Computer-assisted high-performance liquid chromatography method development with applications to the isolation and analysis of phytoplankton pigments. *J. Chromatogr.* A. 910(1): 31–49, doi:10.1016/S0378-434700)00603-4.

Vandermeulen, R.A., A. Mannino, A.R. Neeley, P. J. Werdell, and R. Arnone. 2017. Determining the Optimal Spectral Sampling Frequency and Uncertainty Thresholds for Hyperspectral Remote Sensing of Ocean Color. *Optics Express* 25 (16): 785–97.

Vidussi, F., H. Claustre, B.B. Manca, A. Luchetta, and J.-C. Marty. 2001. Phytoplankton pigment distribution in relation to upper thermocline circulation in the eastern Mediterranean Sea during winter. *J. Geophys. Res.* 106(C9): 19,939–19,956.

Wang, G., Z. Lee, D.R. Mishra, and R. Ma. 2016. Retrieving Absorption Coefficients of Multiple Phytoplankton Pigments from Hyperspectral Remote Sensing Reflectance Measured over Cyanobacteria Bloom Waters, *Limnol. Oceanogr.-Meth.* 14 (7), 425-489, doi:10.1002/lom3.10102.

Wang, G., Z. Lee, and C.B. Mouw. 2018. Concentrations of Multiple Phytoplankton Pigments in the Global Oceans Obtained from Satellite Ocean Color Measurements with MERIS. *Applied Sciences* 8 (2678). doi:10.3390/app8122678.

Werdell, P.J., S.W. Bailey, G.S. Fargion, C. Pietras, K.D. Knobelspiesse, G.C. Feldman, and C.R. McClain. 2003. Unique data repository facilitates ocean color satellite validation, *EOS Trans. AGU* 84, 38, 377.

Werdell, P. J., B.A. Franz, S.W. Bailey, G.C. Feldman, E. Boss, V.E. Brando, M. Dowell, et al. 2013. Generalized Ocean Color Inversion Model for Retrieving Marine Inherent Optical Properties, *App. Optics* 52 (10), 2019–37.

Werdell, P.J., C.S. Roesler, and J.I. Goes. 2014. Discrimination of Phytoplankton Functional Groups Using an Ocean Reflectance Inversion Model, *App. Optics* 53 (22): 4833–49.

Werdell, P. J., L. I.W. McKinna, E. Boss, S.G. Ackleson, S.E. Craig, W.W. Gregg, Z. Lee, et al. 2018. An Overview of Approaches and Challenges for Retrieving Marine Inherent Optical Properties from Ocean Color Remote Sensing. *Progress in Oceanography* 160 (January): 186–212. https://doi.org/10.1016/j.pocean.2018.01.001.

Westberry, T.K., and D.A. Siegel. 2005. An Improved Bio-Optical Model for the Remote Sensing of Trichodesmium Spp. Blooms. *J. Geophys. Res.* 110 (C6): C06012. doi:10.1029/2004JC002517.

Westberry, T.K., E. Boss, and Z. Lee. 2013. Influence of Raman Scattering on Ocean Color Inversion Models, *Appl. Optics* 52 (22), 5552–61.

Wolanin, A., M. Soppa, and A. Bracher. 2016. Investigation of Spectral Band Requirements for Improving Retrievals of Phytoplankton Functional Types, *Rem. Sens.* 8 (10): 871. doi:10.3390/rs8100871.

WoRMS Editorial Board. 2020. World Register of Marine Species. Available from http://www.marinespecies.org at VLIZ. Accessed 2020-03-05. doi:10.14284/170

Wyatt, T. 2014. Margalef's Mandala and Phytoplankton Bloom Strategies. *Deep-Sea Research Part II: Topical Studies in Oceanography* 101: 32–49. https://doi.org/10.1016/j.dsr2.2012.12.006.

Yang, H., G. Lohmann, U. K., M. Ionita, X. Shi, D. Sidorenko, X. Gong, X. Chen, and E.J. Gowan. 2020. Poleward Shift of the Major Ocean Gyres under Global Warming. *Geophysical Research Letters*. https://doi.org/10.1029/2019GL085868.

Xi, H., M. Hieronymi, H. Krasemann, and R. Röttgers. 2017. Phytoplankton Group Identification Using Simulated and In Situ Hyperspectral Remote Sensing Reflectance. *Frontiers in Marine Science* 4 (August): 1–13. https://doi.org/10.3389/fmars.2017.00272.

Xi, H., M. Hieronymi, R. Röttgers, H. Krasemann, and Z. Qiu. 2015. Hyperspectral Differentiation of Phytoplankton Taxonomic Groups: A Comparison between Using Remote Sensing Reflectance and Absorption Spectra. *Remote Sensing* 7 (11): 14781–805. https://doi.org/10.3390/rs71114781.

Yentsch, C.S., and D.A. Phinney. 1989. A Bridge between Ocean Optics and Microbial Ecology. *Limnology and Oceanography* 34 (8): 1694–1705. https://doi.org/10.4319/lo.1989.34.8.1694.

Ye, H., B. Zhang, X. Liao, T. Li, Q. Shen, F. Zhang, J. Zhu, and J. Li. 2019. Gaussian Decomposition and Component Pigment Spectral Analysis of Phytoplankton Absorption Spectra. *Journal of Oceanology and Limnology* 37 (5): 1542–54. https://doi.org/10.1007/s00343-019-8079-z.

Zeng, C., S.Z. Rosengard, W. Burt, M.A. Peña, N. Nemcek, T. Zeng, K.R. Arrigo, and P.D. Tortell. 2018. Optically-derived estimates of phytoplankton size class and taxonomic group biomass in the Eastern Subarctic Pacific Ocean. *Deep Sea Res.* Part I. 136: 107–118, doi:10.1016/j.dsr.2018.04.001.

Zhang, X., L. Hu, and M. He. 2009. Scattering by Pure Seawater: Effect of Salinity, *Opt. Express* 17 (7), 5698–5710.

Zhu, Q., F. Shen, P. Shang, Y. Pan, and M. Li. 2019. Hyperspectral Remote Sensing of Phytoplankton Species Composition Based on Transfer Learning. *Remote Sensing* 11 (17). https://doi.org/10.3390/rs11172001.

APPENDIX: SUPPLEMENTARY INFORMATION TO CHAPTER 5

Image classification

The approach to identifying all imaged diatoms requires multiple steps. The raw data imaged by the IFCB are first processed using a set of codes for image feature extraction: "easyIFCB" (available at https://github.com/OceanOptics/easyIFCB), built on top of the "IFCB_analysis" toolbox (https://github.com/hsosik/ifcb-analysis), using the v4 feature extraction. This prepares the images and metadata for upload to the EcoTaxa web application for image viewing and annotation (https://ecotaxa.obs-vlfr.fr/). On EcoTaxa, images are manually classified into categories representing all identifiable images, as well as "temporary" and general ID categories (e.g., "other" and "detritus"). These manually classified images form a learning set, which was developed and refined over a period of months to eventually contain 14,700 images spread across 58 categories (Table A.1).

Images from individual NAAMES cruises are uploaded to EcoTaxa as separate projects, and the learning set is used to classify all images using a Random Forest machine learning analysis built into the EcoTaxa application. The results of the Random Forest classification are variable in accuracy for different categories, and in general not accurate enough to be used without manual confirmation and correction of the image classification. Therefore, over the course of two years, several people worked intermittently to manually check and correct 2.2 million of the total 4.8 million IFCB images from the four NAAMES cruises. Following the manual correction and classification, there were 123 categories of images, with the majority labeled at the genus level. These were reduced to 18 categories to increase the number of images per category for subsequent deep learning image classification (Table A.2).

To categorize images with greater efficiency we used convolutional neural networks and deep learning. One of the co-authors (E. Culhane) developed an image classification network during summer 2019 that

made use of the 2.2 million classified images for network training. For this work he used the 18-group level image categorization. Images collected using the IFCB are not uniform in their distribution; this presents a major challenge in the development of image classification networks. For example, if all categories at the genus level are included during network development, many do not have sufficient number of images to perform well. However, if these categories are excluded from network training, any images from the excluded categories that are encountered will, by definition, be misclassified. For this reason, as well as to address diatoms as a single phytoplankton group as has been done frequently in the literature, we used the coarse 18-category image classification network.

Table A.1. Image categories used for input to the Random Forest machine learning built into the EcoTaxa application.

Category number	**Category name**	**Number of images**
1	Chaetoceros	425
2	Corethron	212
3	Cylindrotheca	124
4	Guinardia	304
5	Guinardia delicatula	60
6	Hemiaulales	111
7	Membraneis	171
8	Pseudo-nitzschia	66
9	Rhizosolenia	69
10	Thalassionema	26
11	Thalassiosira	125

Table A.1. continued

12	centric	601
13	chain < centric	421
14	pennate	587
15	chain < pennate	356
16	Halosphaera	86
17	Pterosperma	275
18	Pyramimonas	348
19	Ciliophora	347
20	Cladopyxis brachiolata	60
21	Dictyochales	598
22	Dinophyceae	612
23	Torodinium	104
24	Ceratium	218
25	Dinophysis	30
26	Karenia	182
27	Oxytoxum	590
28	Prorocentrum	602
29	Euglenida	484
30	Prymnesiophyceae	151
31	Phaeocystis	82
32	Rhabdosphaeraceae	130

Table A.1. continued

33	Syracosphaerales	45
34	Syracosphaera	172
35	Rhizaria	38
36	Trichodesmium	37
37	bubble	13
38	detritus	265
39	feces	87
40	other < living	603
41	clumps	285
42	othertocheck	638
43	unicellular	728
44	multiple	204
45	part	53
46	t003	223
47	t005	307
48	t007	35
49	t010	22
50	t011	17
51	t012	73
52	t014	832
53	t016	66

Table A.1. continued

54	t017	76
55	t018	35
56	t020	110
57	tempCryptophyceae	617
58	tempflagellates	562

Table A.2. Image categories assigned via manual correction and confirmation of images on EcoTaxa. Columns show the descriptive name (often the same as the EcoTaxa category), and the group assignment for the 18 groups used to train the deep learning image classification network. BQL = Below Quantifiable Limit (7 µm ESD threshold).

EcoTaxa category	**Taxon/descriptive category**	**Group for network training**
Chlorophyta	Chlorophyta	Chlorophyte
Chlorophyceae	Chlorophyceae	Chlorophyte
Halosphaera (Pyramimonadales)	Halosphaera	Chlorophyte
Pterosperma (Pyramimonadales)	Pterosperma	Chlorophyte
Pyramimonas (Pyramimonadales)	Pyramimonas	Chlorophyte
Trichodesmium (Cyanophyceae)	Trichodesmium	Cyanobacterium
Euglenida (Euglenozoa)	Euglenida	Euglenoid
Ciliophora (Alveolata)	Ciliophora	Cilliate
Mesodinium	Mesodinium	Cilliate
Tintinnida (Choreotrichia)	Tintinnida	Cilliate

Table A.2. continued

empty (Tintinnidiidae)	Tintinnidiidae empty	Not living
Dinophyceae (Holodinophyta)	Dinophyceae	Dinoflagellate
Dinophysis (Dinophysaceae Oxyphysaceae)	Dinophysis	Dinoflagellate
Oxyphysaceae	Oxyphysaceae	Dinoflagellate
Ceratium (Ceratiaceae)	Ceratium	Dinoflagellate
Pyrocystis	Pyrocystis	Dinoflagellate
Cochlodinium (Gymnodiniaceae)	Cochlodinium	Dinoflagellate
Gyrodinium (Gymnodiniaceae)	Gyrodinium	Dinoflagellate
Torodinium (Gymnodiniaceae)	Torodinium	Dinoflagellate
Karenia (Kareniaceae)	Karenia	Dinoflagellate
Karenia sp. (Kareniaceae)	Karenia	Dinoflagellate
Warnowia (Warnowiaceae)	Warnowia	Dinoflagellate
Amphidinium (Gymnodiniaceae)	Amphidinium	Dinoflagellate
Oxytoxum (Amphidomataceae)	Oxytoxum	Dinoflagellate
Heterocapsa (Heterocapsaceae)	Heterocapsa	Dinoflagellate
Protoperidinium (Protoperidiniaceae)	Protoperidinium	Dinoflagellate
Scrippsiella (Peridiniaceae)	Scrippsiella	Dinoflagellate
Prorocentrum (Prorocentrum)	Prorocentrum	Dinoflagellate
Rhizaria (Harosa)	Rhizaria	Rhizaria
pennate (Coelodendridae)	Coelodendridae pennate	Rhizaria
Foraminifera (Retaria)	Foraminifera	Rhizaria
Asterionellopsis	Asterionellopsis	Diatom

Table A.2. continued

Hemiaulales (Coscinodiscophycidae)	Hemiaulales	Diatom
Cylindrotheca (Bacillariophyceae)	Cylindrotheca	Diatom
Fragilaria (Bacillariophyceae)	Fragilaria	Diatom
Membraneis (Bacillariophyceae)	Membraneis	Diatom
Nanoneis (Bacillariophyceae)	Nanoneis	Diatom
Navicula	Navicula	Diatom
Nitzschia (Bacillariophyceae)	Nitzschia	Diatom
Pseudo-nitzschia (Bacillariophyceae)	Pseudo-nitzschia	Diatom
Thalassionema (Bacillariophyceae)	Thalassionema	Diatom
Bacteriastrum (Mediophyceae)	Bacteriastrum	Diatom
Chaetoceros (Mediophyceae)	Chaetoceros	Diatom
spore (Chaetoceros)	Chaetoceros spore	Diatom
Ditylum	Ditylum	Diatom
Eucampia (Mediophyceae)	Eucampia	Diatom
Planktoniella (Mediophyceae)	Planktoniella	Diatom
Skeletonema (Mediophyceae)	Skeletonema	Diatom
Thalassiosira (Mediophyceae)	Thalassiosira	Diatom
Thalassiosira weissflogii (Mediophyceae)	Thalassiosira weissflogii	Diatom
Corethron (Corethrids)	Corethron	Diatom
Coscinodiscus (Coscinodiscids)	Coscinodiscus	Diatom
Proboscia	Proboscia	Diatom
Rhizosolenids	Rhizosolenids	Diatom

Table A.2. continued

Guinardia (Rhizosolenids)	Guinardia	Diatom
Guinardia delicatula (Guinardia)	Guinardia delicatula	Diatom
Leptocylindrus	Leptocylindrus	Diatom
Rhizosolenia (Rhizosolenids)	Rhizosolenia	Diatom
centric (Bacillariophyta)	Bacillariophyta centric	Diatom
chain (centric)	Bacillariophyta centric chain	Diatom
pennate (Bacillariophyta)	Bacillariophyta pennate	Diatom
chain (pennate)	Bacillariophyta pennate chain	Diatom
Dinobryon (Chryso clade C)	Dinobryon	Chrysophyte
Dictyochales (Dictyochophyceae)	Dictyochales	Silicoflagellate
Choanoflagellatea	Choanoflagellatea	Other
Crustacea (Arthropoda)	Crustacea	Zoo
Naupilii (Crustacea)	Crustacea	Zoo
Mollusca (Metazoa)	Mollusca	Zoo
Cryptophyta	Cryptophyta	Cryptophyte
Prymnesiophyceae (Haptophyta)	Prymnesiophyceae	Prymnesiophyte
Phaeocystis (Phaeocystaceae)	Phaeocystis	Prymnesiophyte
Prymnesiaceae (Prymnesiales)	Prymnesiaceae	Prymnesiophyte
Syracosphaerales (Prymnesiophyceae)	Syracosphaerales	Prymnesiophyte
Syracosphaera (Syracosphaeraceae)	Syracosphaera	Prymnesiophyte

Table A.2. continued

Scyphosphaera apsteinii (Scyphosphaera)	Scyphosphaera apsteinii	Prymnesiophyte
Rhabdosphaeraceae (Zygodiscales)	Rhabdosphaeraceae	Prymnesiophyte
other (living)	Other	Other
clumps (other)	Clumps	Clumps
multiple (other)	Multiple	Clumps
othertocheck (other)	Other to check	NOT CLASSIFIED
part (other)	Other part	NOT CLASSIFIED
unicellular (other)	BQL	BQL
cyst (unicellular)	Diatom cyst	Diatom
Cladopyxis brachiolata	Cladopyxis brachiolata	Dinoflagellate
not-living	Not living	Not living
artefact (not-living)	Artefact	Artefact
badfocus (artefact)	Bad focus	Artefact
bead (artefact)	Bead	Artefact
bubble (artefact)	Bubble	Artefact
detritus (not-living)	Detritus	Not living
fiber (detritus)	Detritus fiber	Not living
feces (not-living)	Feces	Not living
plastic (not-living)	Plastic	Not living
fiber (platic)	Plastic fiber	Not living
fragment (plastic)	Plastic fragment	Not living
other (plastic)	Plastic other	Not living

Table A.2. continued

t001 (temporary)	T1	Other
t002 (temporary)	Unknown Cocco	Prymnesiophyte
t003 (temporary)	Ophiaster	Prymnesiophyte
t004 (temporary)	T4 Unicellular	Other
t005 (temporary)	Acanthoica	Prymnesiophyte
t006 (temporary)	Lepto_EPI	Other
t007 (temporary)	Cocco-cone	Prymnesiophyte
t008 (temporary)	T8 Unicellular	Other
t009 (temporary)	Quatro-prymno	Prymnesiophyte
t010 (temporary)	Prymnesio-like	Prymnesiophyte
t011 (temporary)	Nano-Dino	Other
t012 (temporary)	T12 Prymnesio	Prymnesiophyte
t013 (temporary)	Cocco-matrix	Prymnesiophyte
t014 (temporary)	Degraded	Other
t015 (temporary)	T15 Unicellular	Other
t016 (temporary)	Dividing cells	Other
t017 (temporary)	Nano-Diatom	Diatom
t018 (temporary)	T18	Other
t019 (temporary)	Prymno-clumps	Prymnesiophyte
t020 (temporary)	T20	Other
tempCryptophyceae (temporary)	Cryptophyceae-like	Cryptophyte
tempflagellates (temporary)	Flagellates	Chlorophyte

Detailed description of the network and its development can be found at: https://github.com/emmettFC/selected-projects/tree/master/plankton_vision. Briefly, the approach uses two models: one to separate detritus and "non-living" particles from plankton, and a second to classify the remaining plankton images into the appropriate taxonomic category. It was determined through trial and error that this two-step approach produced improved results over a single model. Both models use multi-input approaches (similar to that of Ellen et al. 2019) with two branches, where there are two different data types for each IFCB image: a numeric vector of parameter values extracted about the image including time of flight, major axis length, minor axis length, biovolume, etc., and an image pixel matrix. The vectors of values are fed through a multi-layer perceptron network (MLP), and the image data is input to a convolutional neural network (CNN). The output from these two branches is concatenated into a single vector and then fed through several more MLP-style layers before image predictions are made.

The overall model accuracy for all classes is 86%. However, through a random analysis of model "errors", we discovered that the deep learning network in fact correctly assigns labels to images that were mistakenly mislabeled during the manual image classification. Therefore, the model accuracy is higher than the 86% reported. To quantify this, we randomly sampled 200 "errors" from each category and relabeled the images without knowing the original labels. Bernoulli distributions were used to estimate confidence intervals for the proportion of each type of error that was actually correct. We found that overall, 35% of all "errors" are in fact correct classifications, and therefore the overall model accuracy can be adjusted to between 90-91% at the 99^{th} confidence parameter.

Diatoms are the best performing of all plankton categories in the model; 99% of diatoms are correctly labeled by the binary model (separating plankton from detritus/non-living particles). Of these, 99% are

correctly labeled as diatoms within the living plankton categories, resulting in 97% of diatoms observed by the instrument being correctly labeled as such. However, 4% of images from the "other" category and small percentages of non-diatom plankton categories are misclassified as diatoms, resulting in 13% of diatoms images that actually belong in different categories. Adjusting for the labeling errors in the manually corrected images described above means that 9.5 and 11.5% of the diatom "errors" are actually correctly labeled diatoms (the model is essentially correcting the human error in manual image classification). The final result is therefore that the data labeled by the model as diatoms is 90% accurate, and that this population represents approximately 97% of the diatoms observed.

IFCB sample aggregation to decrease statistical counting errors

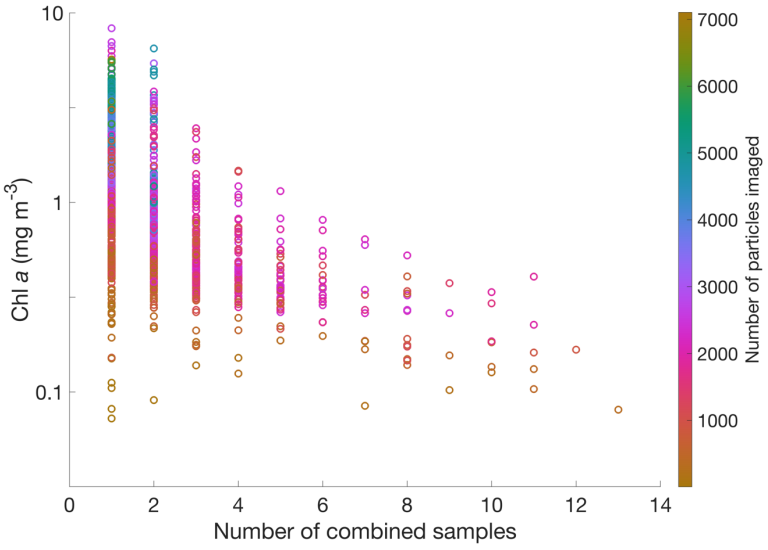

Figure A.1. The number of IFCB samples combined as a function of Chl *a* concentration. Data points are colored by the total number of particles in each combined sample. Number of particles imaged fewer than 2,500 is the result of the 6 hour threshold being exceeded during sample combination before the 2,500 particle threshold is reached.

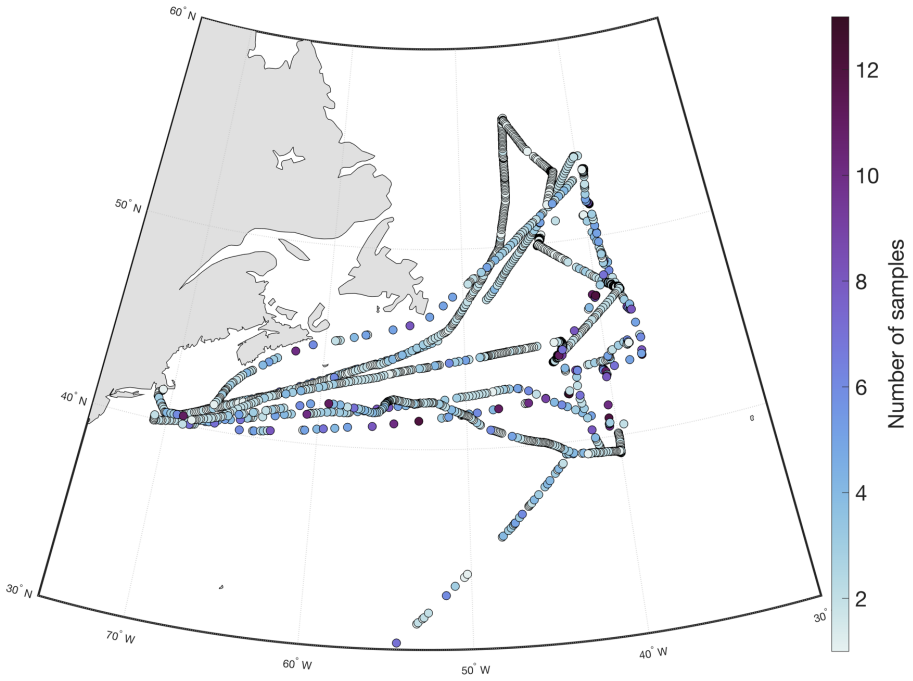

Figure A.2. The number of IFCB samples combined to decrease statistical counting errors in IFCB data. Fewer samples are collected in regions/times of higher plankton biomass/particle counts.

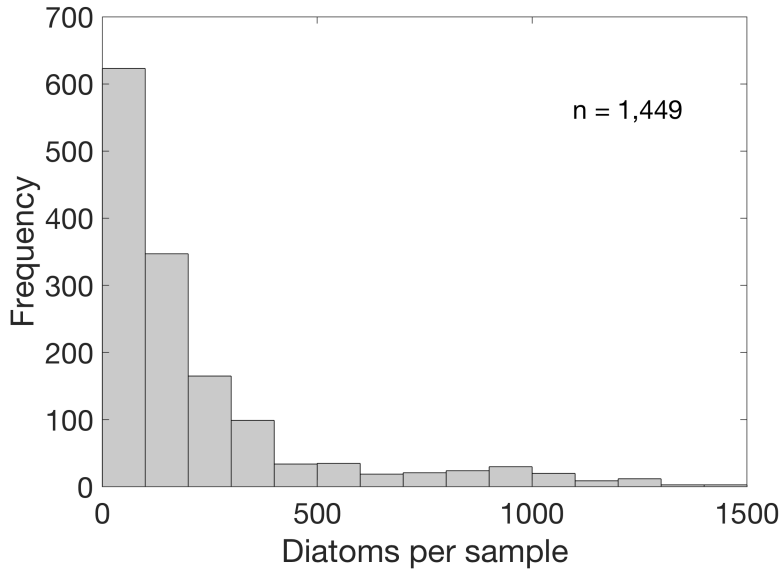

Figure A.3 Number of diatoms per IFCB sample. The minimum value is 20 following removal of samples with < 20. Five samples have > 1500 cells (1924, 2269, 2348, 2383, 3258) and are excluded for ease of viewing.

CHEMTAX Analysis

The CHEMTAX program (Mackey et al. 1996) was run using the MATLAB software, with two different sets of initial pigment ratios and phytoplankton groups: those of Swan et al. (2016), and those of van de Poll et al. (2013) (hereafter denoted as S16 and vdP13, respectively). This results in the contribution of either seven or nine phytoplankton groups (depending on the initiation values used) to Chl *a* for each surface water HPLC data point during the four NAAMES cruises (n = 205). The first set of initial ratios and groups is from the reported results of S16, in which the authors complete a global clustering analysis and determine appropriate pigment ratios and phytoplankton groups for different ocean regions and during each of the four seasons. The North Atlantic Ocean falls within their clusters 5 and 6, defined in S16 as "polar" and "global mesotrophic", respectively. Cluster 5 is defined by the presence of six groups: diatom-1, hapto-6, hapto-8, dino-1, crypto-1, and chloro-1. Cluster 6 is defined by the same six groups plus the addition of cyano-2, for seven groups in total. For the inputs to our CHEMTAX runs, we used the mean values of pigment ratios for clusters 5 and 6 (noting that cyano-2 is included but not as a mean as it is not present in cluster 5). We use these means for the fall values of S16 for NAAMES cruises 1 (November) and 3 (September), and the spring values for NAAMES cruises 2 (May) and 4 (March).

The second set of values is from vdP13 and their associated supplementary information. Their study takes place in the Northeast Atlantic Ocean, and they applied two sets of initial pigment ratios, one for high-light and one for low-light acclimated phytoplankton. As much of our data is from underway ship measurements, we cannot easily diagnose the light history of phytoplankton in the NAAMES dataset. Therefore, we use an average of the high and low light values from vdP13, and the nine phytoplankton groups they report: prasinophytes, dinophytes, cryptophytes, haptophytes_1, haptophytes_2, pelagophytes, *Synechococcus*, *Prochlorococcus*, and diatoms. We use their initial ratio values, as they do not report final values.

For all CHEMTAX runs the settings are as follows:

Maximum number of iterations = 500;

Stop calculation when residual is below = 5.00000000000256E-0003;

Initial step size (inverse) = 10;

Increase step size by a factor of = 1.29999999999927E+0000;

Halt calculation when step is larger than = 1000;

Verbosity = 3;

Vary how many ratios at each step = 5;

Interval between redetermination of what ratios to vary = 5;

Results of the two initializations are noticeably different in multiple phytoplankton groups, but particularly diatoms (Figs. A.4-A.5). Cyanobacteria, cryptophyte, and dinoflagellate contributions are similar for S16 and vdP13. Chlorophytes are prasinophytes (both in phylum chlorophyta) are similar. vdP13 includes pelagophytes, where S16 does not. Both vdP13 and S16 have two types of haptophytes; in combination they match well at times between the two results, and other time the relative contributions of the two haptophyte types and diatoms contribute differently to Chl *a*.

Diatoms differ most noticeably in their contribution to Chl *a* during the first part of the May sampling campaign (sample numbers 95-170). Interestingly, although in general the vdP13 results more closely match the imagery-derived diatom contribution (Figs. A.5 and A.6), the vdP13 results show no diatoms during an extended station occupation in September (sample numbers 187-200), while both imagery and the S16 results show approximately 3-10% diatom contribution to Chl *a*. Overall, vdP13 CHEMTAX inputs show reduced overestimation of diatoms relative to imagery compared to S16 CHEMTAX results (Fig. A.7). The two CHEMTAX outputs also differ in their comparison to the Diagnostic Pigment Analysis (DPA) for diatoms (Hirata et al. 2011), with S16 matching much more closely in the contribution of diatoms to Chl *a* (Fig. A.8).

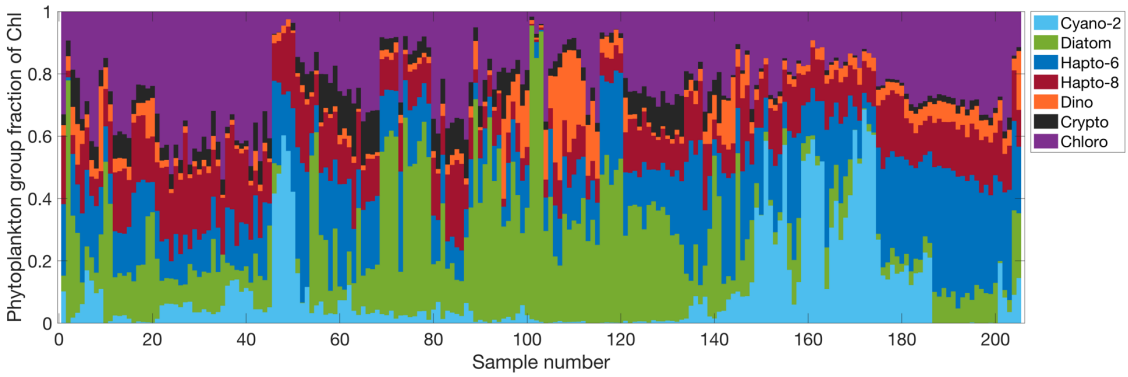

Figure A.4. Phytoplankton group contribution to Chl *a* from CHEMTAX analysis with inputs from Swan et al. 2016.

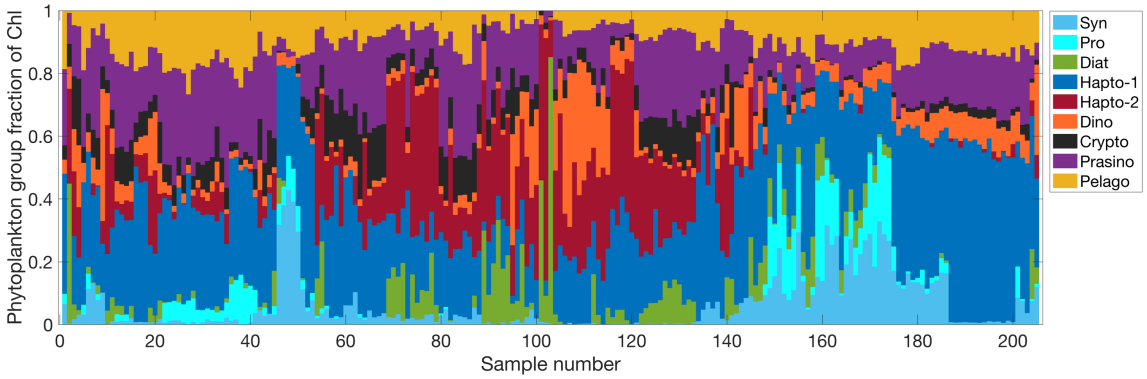

Figure A.5. Phytoplankton group contribution to Chl *a* from CHEMTAX analysis with Inputs from van de Poll et al. 2013.

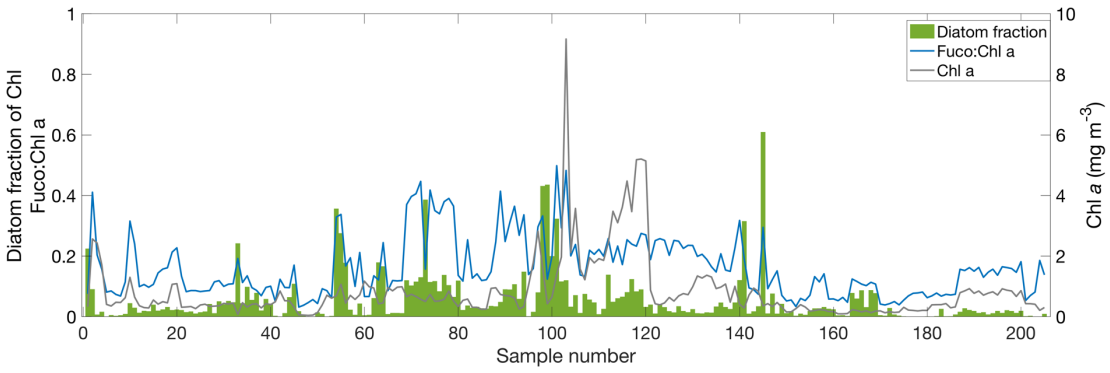

Figure A.6. Diatom contribution to Chl *a* from IFCB imagery. Diatom carbon is converted to Chl using a C:Chl value of 41 (Sathyendranath et al 2009). Diatom Chl is divided by Chl *a* from HPLC to calculate the fractional value. The ratio of fucoxanthin (Fuco) to Chl *a* is shown as the blue line and in left-hand y-axis units. The gray line shows Chl *a* concentration (right-hand y-axis units).

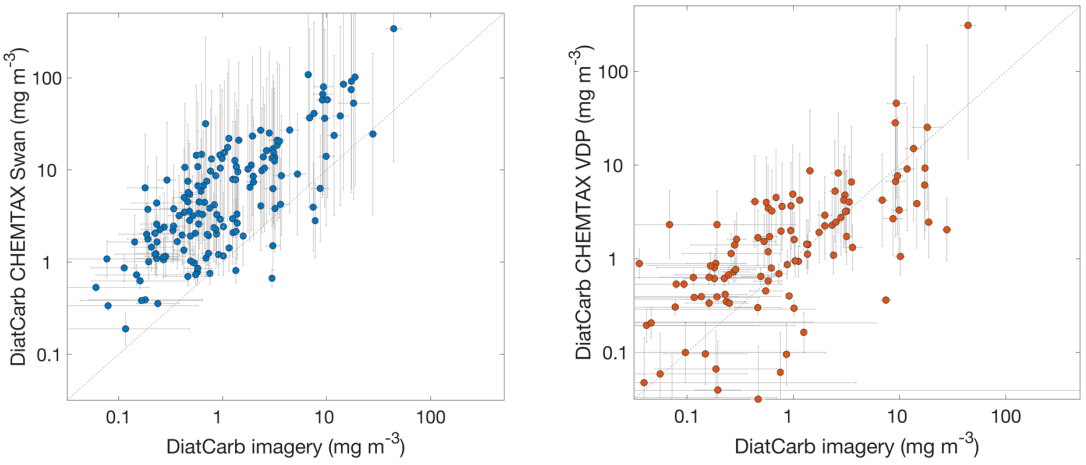

Figure A.7. Diatom carbon from IFCB imagery vs. diatom carbon from CHEMTAX. Left: results calculated using initial ratios and phytoplankton groups from Swan et al. (2016). Error bars in the x direction are from statistical counting errors in imagery data, and error bars in the y direction are calculated from the uncertainty in converting chlorophyll to carbon. Right: same as in the left panel, but with CHEMTAX inputs from van de Poll. (2013).

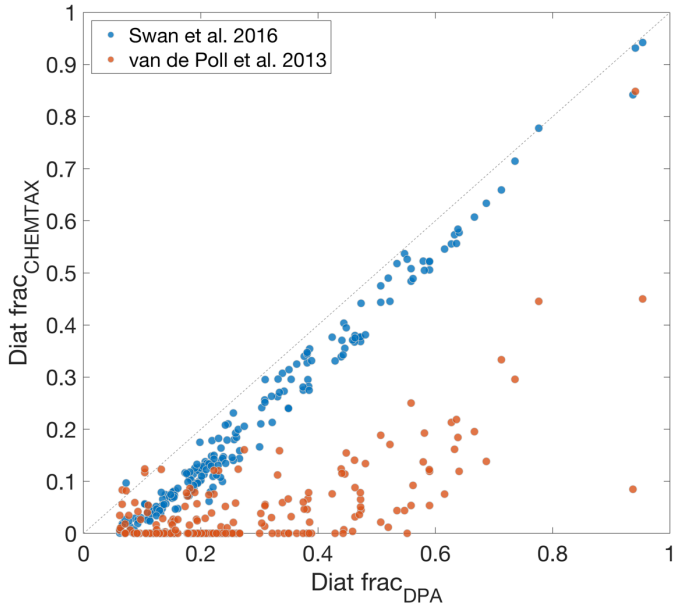

Figure A.8. Diatom Chl *a* from the DPA and from CHEMTAX. The fraction of diatom contribution to Chl *a* calculated as in Hirata et al. (2011) vs. the fraction from CHEMTAX using two different sets of pigment ratio and phytoplankton group inputs.

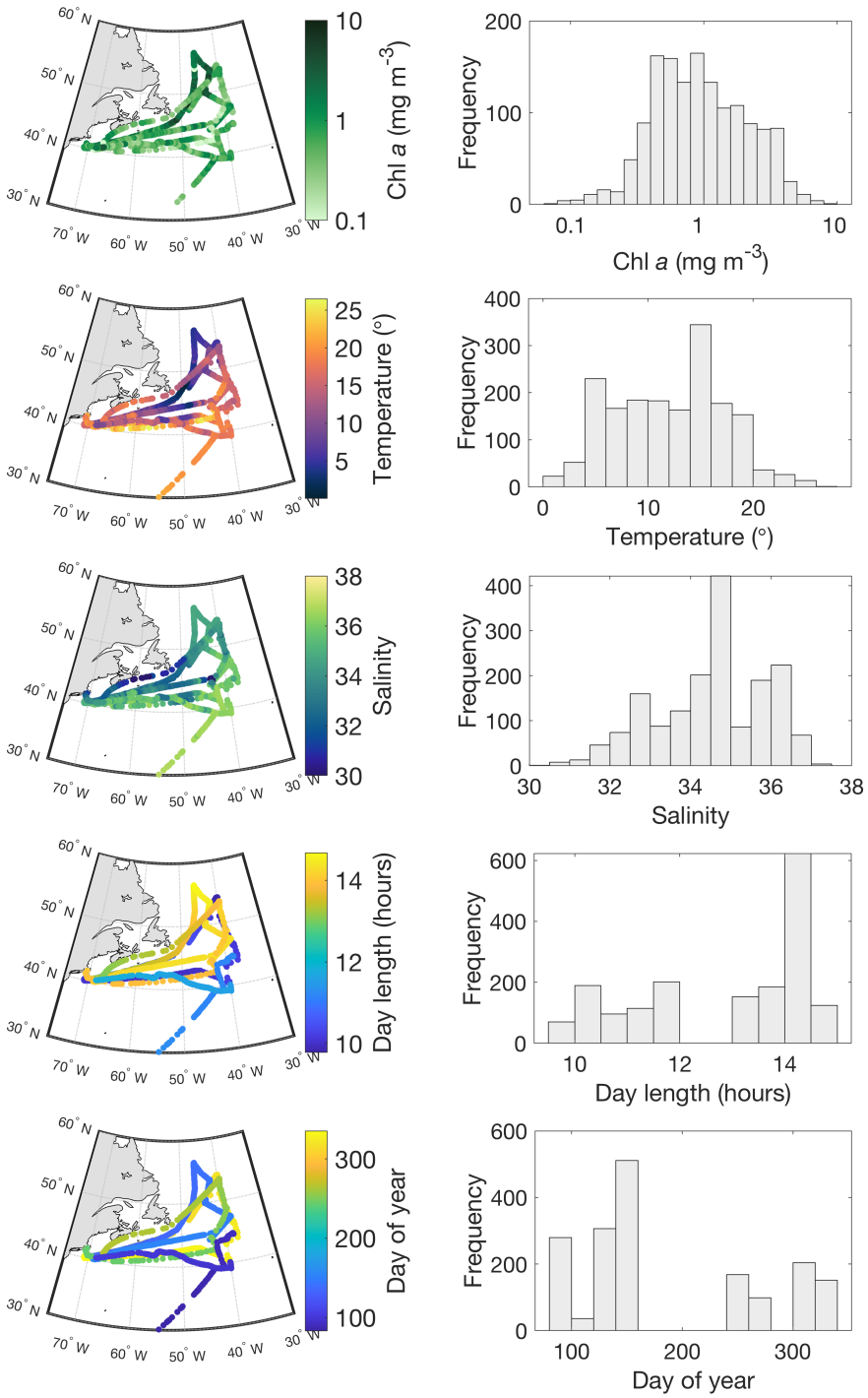

Figure A.9. Locations and frequency distributions of the five input parameters used for neural network training. The four NAAMES cruises provide a wide range of water types, as well as information from across multiple seasons.

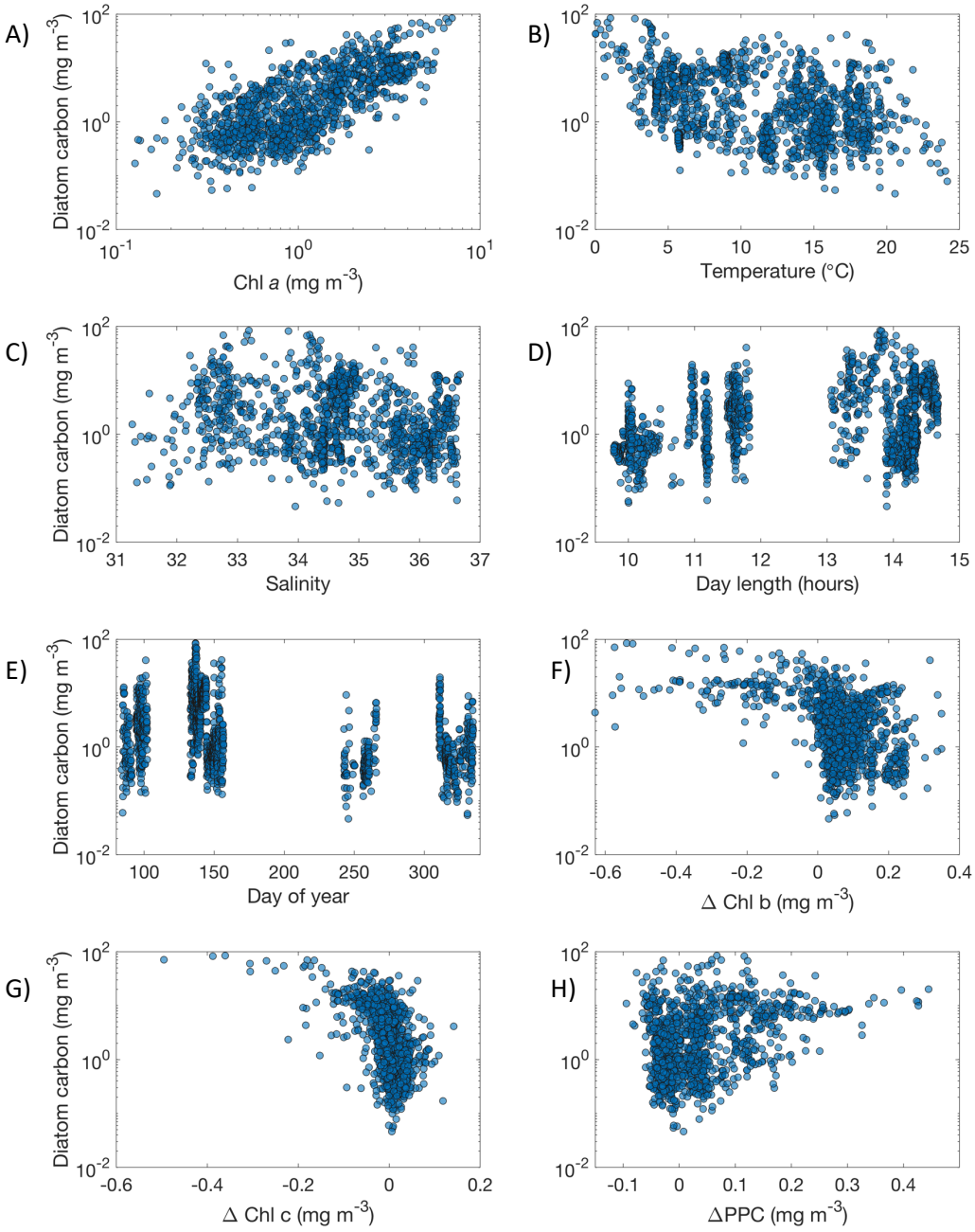

Figure A.10. Diatom carbon from IFCB imagery as a function of environmental parameters and pigment anomalies. A-H: Chl *a*, temperature, salinity, day length in hours, day of year, ΔChl *b*, ΔChl *c*, and ΔPPC. Spearman's rank correlation (ρ) values for the five panels A-H are, in order, 0.7, -0.4, -0.1, 0.3, -0.4, -0.4, -0.5, and 0.3. All correlations are significant with p-values < 0.001.

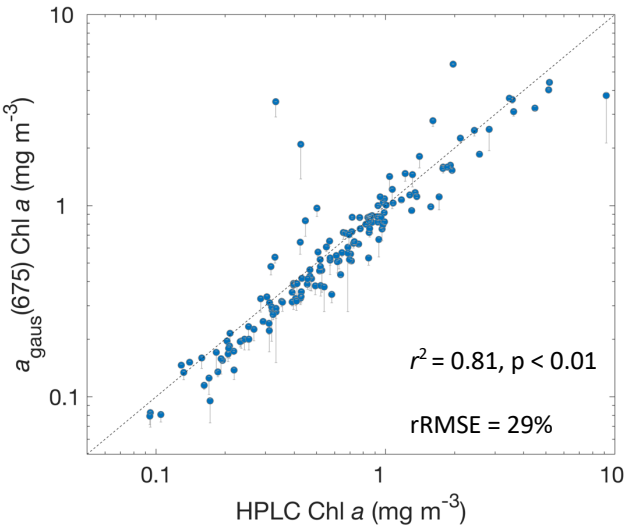

Figure A.11. Chl *a* concentration from HPLC pigment analysis vs. from Gaussian decomposition of $a_p(\lambda)$. HPLC uncertainty is determined through replicate analysis and has a value of 1.41%. $a_{gaus}(675)$ uncertainty is the standard deviation of all values in a +/- ten minute period around the IFCB samples used to compare to the discrete water sample collected for HPLC. Uncertainties in Chl *a* concentrations from $a_p(\lambda)$ decomposition are comparable to those in estimating Chl *a* from the line-height of the red peak in $a_p(\lambda)$ (Boss et al. 2013).

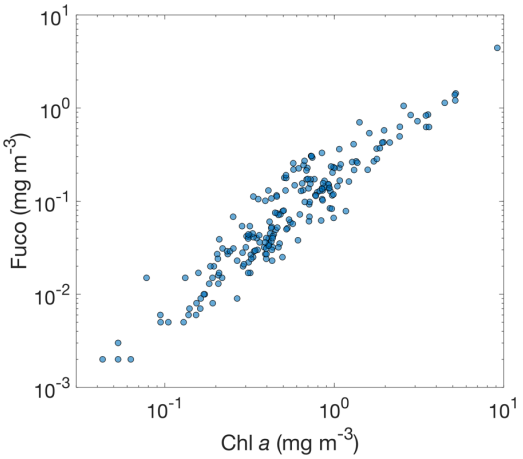

Figure A.12. Chl *a* vs. fucoxanthin as determined from HPLC pigment analysis. N = 205.

BIOGRAPHY OF THE AUTHOR

Alison Chase was raised in Canterbury, New Hampshire and graduated from Belmont High School in Belmont, NH in 2004. She attended Bowdoin College and graduated with a Bachelor of Arts degree in Geology and Environmental Studies in 2009. She spent several years working and returned to Maine and entered the Oceanography graduate program at the University of Maine's School of Marine Sciences in the fall of 2012. After receiving her Master's degree in 2014, Alison continued on as Ph.D. student in the School of Marine Sciences. Following graduation, Alison will move to Seattle, WA for a three-year position as a Washington Research Foundation Postdoctoral Fellow and PACE Science and Applications Team member, working at the Applied Physics Laboratory at the University of Washington. Alison is a candidate for the Doctor of Philosophy degree in Oceanography from the University of Maine in May 2020.